Data Processing Procedures and Methodology for Estimating Trip Distances for the 1995 American Travel Survey (ATS)

Ho-Ling Hwang
Jane Rollow
Center for Transportation Analysis
Energy Division
Oak Ridge National Laboratory

May 2000

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

x

ABSTRACT

The 1995 American Travel Survey (ATS) collected information from approximately 80,000 U.S. households about their long distance travel (one-way trips of 100 miles or more) during the year of 1995. It is the most comprehensive survey of where, why, and how U.S. residents travel since 1977. ATS is a joint effort by the U.S. Department of Transportation (DOT) Bureau of Transportation Statistics (BTS) and the U.S. Department of Commerce Bureau of Census (Census); BTS provided the funding and supervision of the project, and Census selected the samples, conducted interviews, and processed the data. This report documents the technical support for the ATS provided by the Center for Transportation Analysis (CTA) in Oak Ridge National Laboratory (ORNL), which included the estimation of trip distances as well as data quality editing and checking of variables required for the distance calculations.

Approximately 20 percent of the person-trip records or 24 percent of the household trip records were modified by ORNL during the editing process. Each of these records could have one or more fields in it that were edited. These changes or modifications were required in order to improve data quality, perform distance calculations, or to prevent elimination of data. Many of these editing and imputation results were stored in databases or integrated in utility programs and used for subsequent Wave-Cycle data sets. In addition, many of the procedures, utility programs, and databases developed by ORNL during the 1995 ATS processing were also used during the processing of the 1997 Commodity Flow Survey (CFS) data.

ATS data collection began in late spring of 1995 and continued until the spring of 1996. The processing of this data continued into 1997. The ATS public use micro-data files on person and household trips were released by the BTS in 1998, and the associated demographic files were not released until the spring of 1999. See BTS web site *http://www.bts.gov/programs/ats* to obtain ATS data and related information. Most of materials included in this report was based on two interim project reports and a technical memorandum developed by the authors during the duration of the ATS project. The purpose of this report was not only to document the detailed processing procedures for analyzing and using geographic information in survey data, but also to provide our account of lessons learned and recommendations that would be beneficial to other surveys. With the experience gained and the tools and databases developed from this project, we feel that a dramatic reduction in processing time and cost savings can be expected from processing of similar types of surveys in the future.

1. INTRODUCTION

1.1 BACKGROUND

The 1995 American Travel Survey (ATS) collected information from approximately 80,000 U.S. households about their travel during the year of 1995. The ATS is the most comprehensive survey of where, why, and how U. S. residents travel since 1977. ATS is a joint effort by the U.S. Department of Transportation (DOT) Bureau of Transportation Statistics (BTS) and the U.S. Department of Commerce Bureau of Census (Census); BTS provided the funding and supervision of the project and Census selected the samples, conducted interviews, and processed the data. Technical support for the ATS provided by the Center for Transportation Analysis (CTA) in Oak Ridge National Laboratory (ORNL) included the estimation of trip distances as well as data editing and checking of variables required for the distance calculations.

The ATS provides a unique source of information about the use of our nation's passenger transportation system. One of ATS's main objectives is to comply with two requirements of the Intermodal Surface Transportation Efficiency Act of 1991 (ISTEA): (1) to provide "information on the number of people carried in intermodal transportation by relevant classification," and (2) to provide "information on patterns of movement of people carried in intermodal transportation by relevant classification in terms of origin and destination." The ATS provides transportation planners with a useful tool to determine how the transportation systems are being used and who is using them. It also furnishes state/local government and tourism organizations with a comprehensive set of data for transportation and tourism planning.

The ATS is primarily a computer-assisted telephone interview (CATI) survey. A portion of the survey was carried out by personal visit, computer-assisted personal interview (CAPI), particularly for those who could not be reached by telephone. Each ATS sample household was interviewed three to four times during the period of April 1995 to March 1996 (approximately once each quarter) to collect information on each household's travel for the entire year of 1995. Data collected during a particular reference period is referred to as a Cycle. The reference period associated with the first interview Cycle began January 1 to the date of the first interview. Subsequent interviews used the date of the previous interview to the date of the current interview as the reference period. Within each Cycle, households were divided into three representative samples called Waves. A household was assigned to a particular Wave for the duration of the entire survey.

Information collected in the ATS is very comprehensive. It includes travel details for each trip of 100 miles or more (one-way) taken by household members during the calendar year of 1995 as well as demographic data on the household and members of the household. For each trip, the survey gathered information about the destination, overnight stops, mode of transportation, type of lodging, travel dates, trip purpose, reasons for stopping at intermediate locations, and many other travel details. Demographic

information includes household income, race, sex, age, education level, marital status, type of household, type of housing unit, vehicle owned, etc. Information related to travel expenditures and trip distances was not acquired during the interview.

The travel distances for trips collected in the ATS are vital for several reasons. First, an accurate estimate of trip distance was required to determine whether a particular trip was within the scope of the ATS, i.e., 100 miles or more (one-way). Secondly, it is important to be able to associate travel distance with trip characteristics in order to understand travel behaviors among various socioeconomic or ethnic groups. Finally, reliable distance estimates are crucial in the calculation of vehicle-miles traveled (VMT) and person-miles traveled (PMT) on the U. S. transportation systems. These estimates of VMT and PMT play important roles in the distribution of transportation funds to the states.

In a pretest version of the ATS questionnaire, survey respondents were asked to provide an estimate of the one-way distance from their residence to the farthest point of their trips. Data from this pretest survey suggested that respondents were often unable to provide accurate estimates of the distance they traveled. This question, along with several others, was dropped from the actual questionnaire used in the ATS in an effort to reduce the survey respondent's burden during the interview.

1.2 SPECIAL SECURITY CONSIDERATIONS

In order to enable ORNL staff to recover as much data as possible and to provide accurate distance estimates, it was necessary for ORNL to work with the micro data collected from the ATS survey. Because the data is protected by U.S. Code Title 13, which requires all necessary precautions and safeguards be taken to protect the confidentiality of Census data, ORNL became a Certified Secured Data Processing Facility. Several facility characteristics of ORNL were especially important in obtaining this certification. ORNL is a federally-owned, contract-operated nuclear facility designed to provide contractual research and support to the Department of Energy (DOE), the Department of Defense (DoD), and other government agencies. Systems used to process ATS data reside in buildings within the same fenced, 24-hour guarded access-controlled portion of the ORNL facility of Bethel Valley Road. All personnel are subject to background investigations commensurate with the sensitivity of the data to be handled, and project team members who have access to ATS data hold a security clearance or have had satisfactory background investigations. Facility emergency personnel, fire captains, and plant shift supervisors, as well as key control personnel, have had background investigations and are cleared to the highest level.

Special security procedures were also implemented and followed by ORNL staff for the transfer, processing, and physical storage of all ATS data. All ORNL staff who worked on the ATS project became Bureau of Census Special Sworn Employees (SSE). As such, they were initially trained in security measures for handling Title 13 sensitive data and were required to follow a security plan that was submitted and approved by Census. Project personnel were required to sign a sworn affidavit

concerning the requirement not to disclose Title 13 data and were also given frequent security refresher training courses.

Data was transferred from Census to ORNL through Federal Express mail that was hand-carried or picked up so that it did not go through plant courier mail. The system used to process ATS data initially consisted of two workstations, a UNIX-based Sun SparcStation with two X-terminals connected to a local area network, and an Intel-processor Gateway 2000. These systems resided in separate offices with key-controlled access inside the fenced, 24-hour guarded portion of ORNL.

In the fall of 1997, the ATS team and the ATS computer equipment moved to a new building also within the same fenced, 24-hour guarded area of ORNL. This new facility has a special laboratory space with a scramble, cipher lock and a limited (grand master) key core. In addition, several rooms were designated as special access rooms for ATS and use an ORNL grand master key series available only to SSE and emergency personnel.

All processing of ATS data was done on these systems and on systems that had removable external hard drives within these authorized rooms. Except for the Sun workstations, which will be secured until they are disposed of, data was viewed, edited, and processed on removable data cartridges, never on an internal hard disk. All personal computers (PCs) had to be disconnected from the network whenever an ATS data cartridge was inserted into the system. Any printed material was properly labeled and kept in locked files, or safes, until it was shredded. Data cartridges were stored in two fire-proof safes in key-controlled areas.

1.3 GEOCODING OF PLACE NAMES

1.3.1 Geocoding for Domestic Locations

Accurate distance estimates require that the correct geographic locations (i.e., latitude and longitude) of the origin, destination, and intermediate stops be identified. Ideally, trip distance measured at the lowest possible geographic level such as address-to-address (i.e., door-to-door) would be the most desirable. This type of measurement, however, would require tremendous effort in collecting precise location information for each trip from the survey respondent. It would also call for enormous address matching and geographic coding activities. Following the lead of the 1993 Commodity Flow Survey (CFS), geographic coding for the 1995 ATS trip data was set at the zip code level. That is, the length of a trip segment was determined based on the distance measured between two zip codes where the end points of the segment were located. Because a zip code typically represents an area not a point, a certain point-location within the given zip code was selected so that a coordinate could be assigned. Consequently, the distance of a trip segment could be calculated mathematically from the coordinates of its two end points.

Version 3.2 of the 5-Digit Zip Code Inventory File, a product of Geographic Data Technology (GDT), Inc., was used to geographically code the domestic place names in the 1995 ATS trip files. This version of the zip code file contains the United States Postal Service (USPS) 5-digit zip codes as of February, 1995. Geographic coordinates (i.e., latitude and longitude) of a zip code centroid, which represents the balance point of the polygon formed by its boundaries, were calculated by GDT and were provided in this zip code inventory file.

Since the population is seldom distributed homogeneously within a zip code, it was determined that a population weighted centroid instead of the geographic centroid of a zip code would provide more accurate measurement in trip distances. Public Law 94-171 Data on CD-ROM (PL-94) from the Bureau of Census was utilized to estimate geographic coordinates of the population weighted centroid for each zip code. ORNL's ATS research team selected the block- level geographic entities from the PL-94 data file. Calculation of the geographic coordinates (latitude and longitude) of a population weighted zip code centroid was based on geographic coordinates of the block centroid, weighted by associated block level populations, of all blocks within the given zip code. Effort was made to ensure that every population weighted zip code centroid fell within its boundary (i.e., an internal point). The method used by the Census to relocate a centroid from outside its boundary, or from falling in a water area, was adapted by ORNL in this work (see Appendix A of the PL-94 CD-ROM documentation).

With the exception of point zip codes (e.g., post office box or specific building location), a population weighted zip code centroid for each of the zip codes contained in the GDT inventory file was calculated by ORNL. These new coordinates along with the zip code level populations were added to form an updated zip code inventory database. This updated zip code inventory database was then used as the base for the geographic coding of domestic place names in the ATS trip data.

For ATS trip files, all zip codes for domestic locations, on which calculation of the trip distances was based, were assigned by a systematic Autocoder at the Census. A list of zip codes along with their associated place names was extracted from the GDT zip code inventory file by ORNL and provided to the Census at the beginning of the ATS project. This information was used by the Census to update their existing Autocoder system. Locations not matched with zip codes by the Autocoder were sent to clerical staff (also Census) for manual zip code matching. Many locations left un-matched were identified as "uncodable" (i.e., geocoding couldn't be performed by the Census). These uncodables also included all foreign locations as well as airports, stations, and ports in the United States.

1.3.2 Geocoding for Foreign Locations

ATS captured many long distance trips that involved traveling to destinations in a foreign land and/or took intermediate stops during the trip in foreign locations. The 5-digit zip codes were not used to geographically code these foreign place names for obvious reasons. Geographic coordinates for these foreign place names were coded by ORNL using a database created from several sources. The

database was built, initially, based on information extracted from the Digital Chart of the World.[1] Additional foreign place names were appended to the database as new entries (i.e., names not already in the database) were encountered during the processing of ATS trip data. These include both correctly spelled location names, full or abbreviated, and many possible ways of misspelling certain place names.

Geographic coordinates of foreign places for those not found in the Digital Chart of the World were entered manually to the database. References used to determine these coordinates were obtained from *The Times Atlas of the World*[2]. *The Times Atlas of the World* provides latitude and longitude for all place names found on the maps in its index section. *The MacMillan World Atlas*[3] was used as a supplement to identify additional foreign locations not found in the Digital Chart of the World and *The Times Atlas of the World*. Geographic coordinates for these place names were estimated from the latitude-longitude grids on the associated maps. For a place name that could not be identified, or was missing, geographic coordinates of the centroid of the country where the place is located were used. No attempt was made to separate geographical coding of foreign airports, train stations, or water ports, however. All of them were defaulted to the place names.

1.4 PURPOSE OF THE ATS DATA PROCESSING PROCEDURES

The CTA was tasked by the BTS to (1) review 1995 ATS trip data in order to improve the quality of the data and (2) provide mode-specific distance estimates for each of the ATS trips. For this survey, BTS was interested in collecting long distance trips that were 100 miles or more one-way (either from home to the destination or from the destination to home). In order to reduce the possibility of lost trips that should have been within the scope but were not reported due to the respondent underestimating trip length, the ATS respondents were asked to report all trips that were 75 miles or more (one-way). After the distance calculations, trips that were under 100 miles one-way were marked for deletion. The distance estimates, therefore, were crucial in determining whether or not a trip was within the scope of ATS.

ATS data was sent from the Census to ORNL in a flow-based pattern, identified by Waves and Cycles. Each Cycle was associated approximately with a quarter of the calendar year, and each Wave

---

[1]Digital Chart of the World, Edition 1. A digital database of the world developed by the United States Defense Mapping Agency (DMA) with the cooperation of Australia, Canada, and the United Kingdom.

[2]*The Times Atlas of the World*, Ninth Comprehensive Edition. Times Books, a Division of Random House, Inc. New York and Canada. 1994.

[3]*The MacMillan World Atlas*. A Simon and Schuster MacMillan Company, New York, NY, 1996.

within a Cycle was associated approximately with a particular interview month. Table 1 gives an overview of the reference period (i.e., recall period) for each Wave and Cycle and the work load for the data processing activity at ORNL. Since each trip was coded as segments (i.e., from one point to the next point along a given trip), the number of trip segments reflects the number of records requiring distance calculations.

Table 1. ATS Data Overview

	No. of Trip Segments (Records)	No. of Trips	No. of Households	Interview Month	Reference Period
Cycle 1 Wave 1	106,566	33,668	12,057	May	Jan. - May
Wave 2	124,426	38,777	12,777	June	Jan. - June
Wave 3	142,447	44,193	13,523	July	Jan.-July
Cycle 2 Wave 1	100,050	31,269	11,751	August	May - Aug.
Wave 2	105,723	32,535	12,001	September	June - Sept.
Wave 3	94,468	29,779	11,481	October	July-Oct.
Cycle 3 Wave 1	84,964	26,792	10,603	November	Aug. - Nov.
Wave 2	73,552	23,281	9,755	December	Sept. - Dec.
Wave 3	61,089	20,126	8,918	Jan./Feb.	Oct.- Dec.
Cycle 4 Wave 1	41,955	14,029	7,294	February	Nov. - Dec
Wave 2	21,978	7,805	4,813	March	Dec.
Total	957,218	302,254			

Travel data on all trips collected during each Wave and Cycle of the ATS interviews were compiled into SAS data sets by the Census. After the initial assignment of zip codes by the Census, all trip files were transmitted via computer tapes to ORNL for mode-distance calculations. Because the accuracy of the geocoding was essential for the precision of distance estimates, extensive checking and editing procedures were developed by ORNL to identify zip codes for those "uncodable" domestic locations and to assign latitudes and longitudes for airports, stations, and all foreign locations. Checking for other missing or incorrect zip codes was also performed during ORNL's data processing. If the zip code of a domestic place name was not provided by the Census, ORNL would first attempt to assign its zip code by matching this place name to the zip code inventory database. Those that failed to be matched with valid place names in the zip code inventory database were then matched with the most recent

version of the ZIPFIXED database (see Section 7.1) to locate a zip code. Often, multiple zip codes could be associated with a common place name. In such cases, the most populated zip code (i.e., the zip code with the highest number in population) for that given place name would be used to provide geographic coordinates needed by the distance calculation.

Note that the zip code of an airport (or an Amtrak station), rather than the zip code of the city where the airport (or Amtrak station) is located, was used whenever possible when filling in missing zip code variables for the ATS trip records. The distance calculations, however, were based on the geographic coordinates of these airports or stations and not on the centroid of their zip codes.

Common problems encountered by ORNL during the ATS data processing include the following:

1.   Interviewees frequently reported destinations as the name of a resort, state park, amusement park, school, etc., rather than a particular city and state.
2.   Zip codes were often incorrectly assigned or un-assigned due to incorrect coding of the state abbreviations in the trip records. Common mistakes included confusion in the 2-character state code for Alaska, Alabama, Arizona, and Arkansas, as well as among Mississippi, Michigan, Minnesota, and Missouri.
3.   Misspelled city names also resulted in difficulty in assigning zip codes; this problem was most evident in domestic location names such as Albuquerque, New Mexico, which had over 51 misspellings for the city name (e.g., Aberquerque, Abuquerque, Abequerque, Abaquerque, etc.). Misspelled foreign place names were especially common, where ORNL matched the place names with the latitude and longitude so that distance could be calculated. Table 2 shows examples of various spellings and codings of location names for Vancouver, Canada.
4.   Various abbreviations of airport names caused difficulty in matching names with latitude and longitude. For example, John F. Kennedy International Airport was abbreviated many different ways (e.g., JFK, J.F.K. J F K, JFK INTL, Kennedy, etc.).

Discussions of these and other types of data problems as well as procedures used to fix them are presented in more detail in Section 3 of this report.

Once the initial geocoding and editing was completed, ORNL calculated trip distances for each trip. The methodology used in ORNL's distance calculation is mode specific. ORNL's highway network database was used to route highway trips using a shortest-path algorithm. Train trips were routed on the Amtrak network database with a similar algorithm. Great circle distances were used as the distance between two airports or two water ports. A more detailed description on methodologies used for the ATS distance calculation can be found in Section 5 of this document.

Originally, it was both BTS's and ORNL's intention to recover as many trip records as possible. Thus, during the processing of Wave-one-Cycle-one (W1C1) data, ORNL utilized many tools to locate place names and zip codes and to resolve problems with particular cases. These tools included hard

Table 2. Sample Entries for Vancouver, Canada*

Destination City	Destination State	Foreign Country
Vamcouver	CD**	
Vam Couver	FC	CD
Van Couvee	CD	
Van Couver	CD	
Van Cover	FC	British Columbia
Van Cover Island	FC	British Colu
Van Courver	FC	Canada
Vancoover	FC	British Columbia
VanCouver	FC	Canada
VanCouvee	FC	Canada
VanCouverh	FC	

 * Sample of a total of 72 combination entries of destination city, destination state, and foreign country for
   Vancouver, Canada.
** By Census's coding definitions, "MX" was to be used for coding locations in Mexico, and "CD" was to be used
   for coding locations in Canada; "FC" was to be used for all other foreign countries.

copy atlases, phone directory CD-ROMs, geographic software products, Internet searches, and other
reference materials. A draft data processing procedure was assembled during the first few months of
the W1C1 data processing period, and the final report is included in later sections of this document.
This data processing procedure report served as a living document to guide the data processing
activities in subsequent Waves and Cycles. It also specified the database structures and record layouts
for files being transmitted between ORNL and the Census. The data processing procedures and report
were updated during the ATS operation as new problems and their solutions were identified.

1.5 ORGANIZATION OF THE REPORT

This report describes the nature and scope of the 1995 ATS and ORNL's role in data editing and
processing of the ATS data. The rest of this report is organized as follows. First, Section 2 describes
the operational flow of the ATS data processing activities. Section 3 describes more detailed
procedures that were followed by ORNL for data screening and pre-processing. After the completion
of the first Cycle of data, ORNL was asked to modify the original processing procedures to reduce

processing time and produce results by an earlier schedule. These modified procedures, called "speedy processing" procedures are discussed in Section 4. Section 5 describes the techniques used to estimate origin-destination (O-D) distances for trip segments by mode: highway, air, train, and water. Quality Control (QC) and Quality Assurance (QA) checks on data and results as well as methods for handling problem records are discussed in Section 6. Section 7 discusses the implications of using the detailed data processing procedures used on the Cycle 1 data versus the "speedy processing" approach. The final section, Section 8, presents ORNL's lessons learned and recommendations that could be useful for future ATSs and similar surveys. Most of the materials presented in this report were produced as project interim reports or technical memoranda by the authors during the ATS data checking, editing, and distance estimation processes. The first report, "ATS Data Processing Procedures," was a draft report delivered to BTS in March 1996, while the second draft report "Impact of ORNL Quality Assurance on the 1995 ATS Trip Data" was sent to BTS in December of 1997. In January 2000 a technical memorandum, "Suggestions for ATS 2000" was submitted to BTS.

2. DATA PROCESSING FLOW

The measurement of travel distance in the ATS is a crucial element for two major reasons. It not only provides a base for selecting in-scope trips (i.e., 100 miles or more one-way) but also produces an estimate of total miles traveled from origin to destination and return for each trip. The calculation of distance for a trip segment relies on the location information of the two end points (i.e., origin and destination). Consequently, the accuracy of this measurement, besides the impact from the calculation method itself, depends on accurate geographic locations of the trip origin and destination.

For this reason, the ATS data processing activities at ORNL included several stages. A detailed description of the operational flow for the ATS data processing is presented in this section. Specific procedures or methodologies used in each of these stages are further elaborated in later sections.

2.1 OVERVIEW OF THE OPERATIONAL FLOW

Data processing operation at ORNL included three major stages: pre-processing, distance calculation, and post-processing. Each of these components, or data processing stages, contained several steps. For instance, pre-processing included activities such as data screening, automatic error checking and editing, as well as categorization of trips for specific operations. These stages and their associated steps are discussed in more detail in Section 3.

Figure 1 shows the operational flow for the ATS data processing activities conducted at ORNL. For each of the ATS Wave-Cycle specific data sets received from the Census, ORNL's data processing operation began with database conversion, data screening and pre-processing, distance calculation, post-processing, and finally ended with sending the result file back to the Census. This operation was reinitiated and repeated as each Wave-Cycle data file was received from the Census. Therefore, at any given time, several Wave-Cycle data files could be processed simultaneously at different stages in ORNL.

2.1.1 Data Transmittals Between the Census and ORNL

For each trip collected during the ATS, Census assigned zip codes to the reported origins and destinations by using computer (i.e., auto-coder) as well as clerical processes. Place names that could not be coded by the Census were left blank for ORNL's further review. After this zip code matching operation, ATS data was converted to SAS transport format and copied onto computer tapes by the Census. These data tapes were transmitted to ORNL via Federal Express. The data transfer process followed a flow-based pattern of Waves and Cycles as described in Section 1.2. To ensure the maximum security of these data, data tapes were held at the Federal Express office as they were

Receive Data Tape from Census

Convert Data to PC Database Files

Generate Quick Summary

(statistics on no. of trips, trip types, modal shares, no. of stops per trip, etc.)

Separate Data into Various Files for Editing/Checking

| Commuter Trips | Mover Trips | Add-Stop Trips | Trips w/ Uncodables | Trips w/ HI to HI | Waterway Trips | No-stop Trips | Other Trips |

Checking/Editing

No

OK?

Yes

No

OK?

Yes

Manual Review

Water

Computer-Aided Systematic Review & Checking

Air/Train

Separate Trips by Mode(s) Used

Highway Only

Distance Calculation

No

OK?

Yes

Generate New Records for Cases w/ Border Crossings or Mode Changes

Compile Output File for the Census

Convert File to SAS and Copy to Tape for Delivery to Census

Figure 1. Operational Work Flow for Processing ATS Trip Data

12

received and hand-carried by authorized ORNL project staff to the Census Secured Data Processing facility in ORNL.

Procedures for returning result data tapes from ORNL to the Census were similar. Output data files were converted to SAS transport format and copied onto computer tapes. Authorized ORNL project staff then hand-carried the tapes to a local Federal Express office for delivery to the Census.

2.1.2 Definition of Trip Segment Type (i.e., Record Type)

All ATS trips were coded as segment records by the Census. A record type variable was provided in each record to distinguish what the segment represented. All segments within a given trip had the same identification variables (case ID, person ID, and trip ID) which serve as the linkage for reconstructing the trip. All trips, with the exception of moving trips and commuter trips, were defined and coded as round trips. Thirteen record types were used in the ATS Wave-Cycle trip file. A brief description of each of these record types is presented here.

A. Record types used in the coding of regular long distance trips

Type 1:
This is the main record which holds all trip characteristics associated with the 'to' direction (i.e., out-bound trip away from home) of a trip. This trip record contained the respondent's residence at the time of travel as the origin and the final destination (i.e., the farthest point traveled on a trip) as the end of the trip segment. Typically, this origin is where the respondent resided at the time of survey, unless he/she had moved from another location during 1995 but prior to the interview.

Type 2:
For travel that utilized a transportation facility to begin the journey, such as an airport, a pier, a bus terminal, or a train station, a record Type 2 was included. This record type provided information for the access link of the trip. This trip segment used the origin of a trip as its origin and the location of this transportation facility as the destination. Geographic detail regarding the facility, if available, was also provided in this record.

Type 3:
This trip segment went from the location of a transportation facility, as given in the Type 2 record, to the final destination of the trip. A Type 3 record existed only if a Type 2 record was present in the same trip. Furthermore, there must not be any intermediate stops reported for the out-bound travel (i.e., the 'to' direction of a trip). In other words, record Type 3 was used exclusively in coding travel that used transportation facilities to begin the trip but made no other stops before reaching the final destination.

13

Type 4:

This type of record was used to identify segments associated with stops along a journey from home to the final destination (i.e., the 'going' trip). A Type 4 segment typically followed a Type 1, Type 2, or another Type 4 record. It started from a location identified as the destination in the preceding segment and ended at the location of an intermediate stop along the travel path. For each given intermediate stop on the trip path, two segments would be defined. The first segment would use the location of this given stop as its destination, followed by another that began with this location as its origin. All attributes of the given stop (e.g., reason for stopping, lodging type, etc.) were contained in the first segment record (i.e., the one that ended at the stop concerned). The number of Type 4 record segments for a given trip, therefore, is equal to the number of intermediate stops along the path of the 'going' trip, plus one.

Type 5:

Record Type 5 was used exclusively for overnight side trips taken by travelers while staying at their final destinations. Multiple side trips were allowed for each trip. The coding was similar to that of the Type 4 record, and information associated with each of the side trip locations was also included.

Type 6:

This type of record was used to identify segments associated with stops along the travel path from the final destination back to home (i.e., the return stop). Similar to the Type 4 record, this record started from a location identified as the destination in a preceding segment record and ended at the location of the intermediate stop reported. All attributes of the given stop (e.g., reason for stopping, lodging type, etc.) were contained in the segment record that ended at this stop. As in the Type 4 record, the number of Type 6 segment records is equal to the number of return stops plus one.

Type 7:

This is the main record which holds all trip characteristics associated with the 'from' direction (i.e., in-bound trip returning to home) of a trip. This trip record contained the final destination of a trip as the origin and the respondent's residence at the time of travel (i.e., home) as the destination.

Type 8:

This record type holds additional information on the added stops for the given trip. Occasionally during an interview session, a survey respondent might provide supplemental information on additional intermediate stops for a trip that has already been reported. Census's survey instrument did not insert this additional information back to the trip record. Instead, these data were appended to the file as "add-stops." All add-stop records were coded with associated trip identification variables and assigned a record

Type 8. In order to make the best use of all data collected from the survey, these 'add-stop' records were reviewed by ORNL and manually inserted into the trip.   Re-sequencing of trip segments was conducted whenever necessary. The practice, unfortunately, required substantial time and tremendous effort.   In an effort to speed up the ATS data processing operation, this activity was abandoned after the completion of Cycle one data.

B. Record type used in the coding of long distance commuting trips

Type 9:
This trip record contained the respondent's residence at the time of travel as its origin and his/her work place as its destination. Typically, the origin is where the respondent lived at the time of survey, unless he/she had moved from another location during 1995 but prior to the interview.

C. Record type used in the coding of long distance moving trips

Type 10:
This is the main record that holds all trip characteristics associated with the moving trip. The trip record contained the respondent's old residence location as the origin and his/her current residence location as the destination of the trip segment.

Type 11:
This record type was the same as the Type 2 record except it was used exclusively for moving trips.

Type 12:
This record type was the same as the Type 3 record except it was used exclusively for moving trips.

Type 13:
This record type was the same as the Type 4 record except it was used exclusively for moving trips. Multiple Type 13 segments were possible. No Type 12 record would be used concurrently with a Type 13.

Figure 2 provides several diagrams to illustrate the use of these record types in expressing various types of trips. For example, if a respondent lived in Detroit, Michigan, and took a trip by car to Tampa, Florida, with a 'to' stop in Atlanta, Georgia, and a 'return' stop in Lexington, Kentucky, the trip would contain the following record types.

1. Mover trips
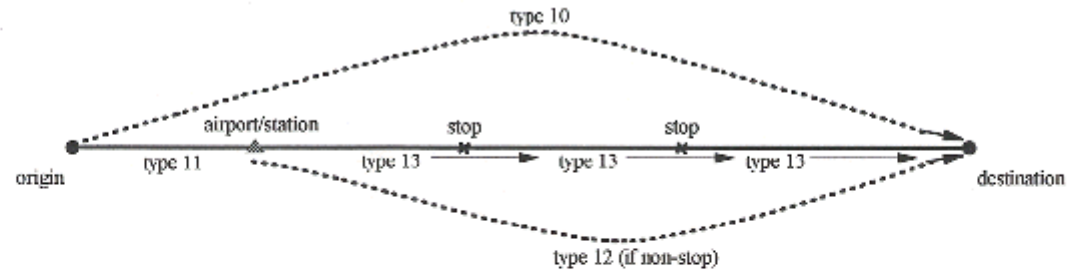
2. Commuter trips
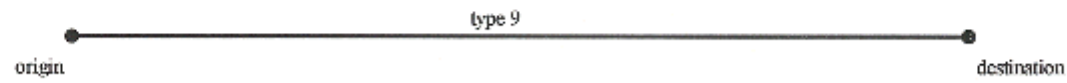
3. Driving trips
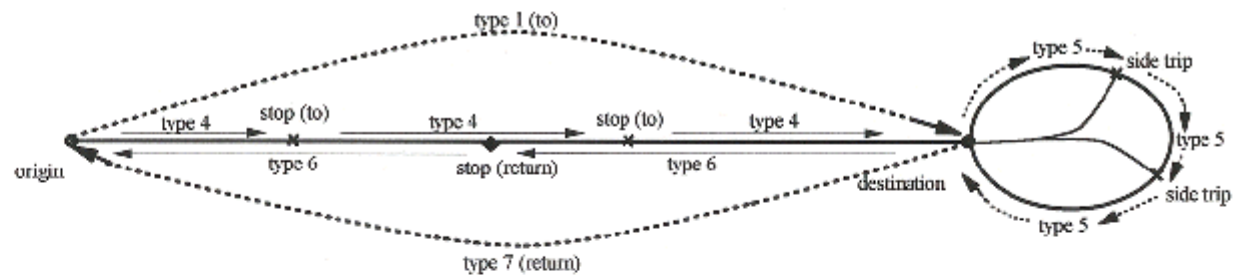
4. Air/bus/train/water trips
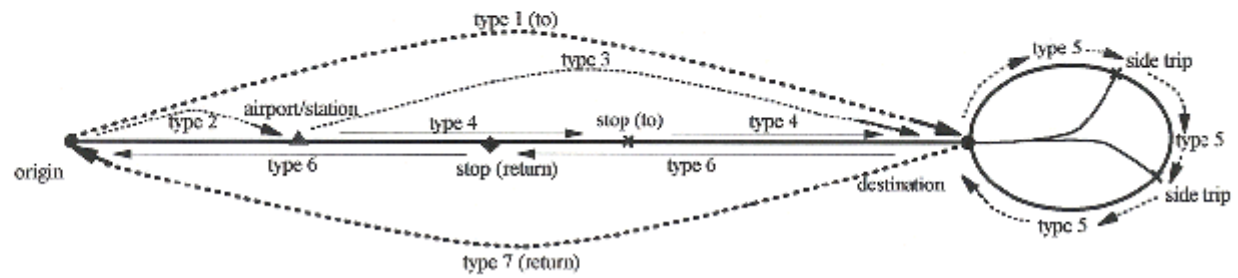
Figure 2. Trip Diagrams for the ATS

16

A. Going 'to' trip

- A Type 1 trip record with an origin of Detroit, Michigan, and a destination of Tampa, Florida (main 'to' record).
- A Type 4 trip record for the 'to' stop segment with an origin of Detroit, Michigan, and a destination of Atlanta, Georgia.
- A Type 4 record for the 'to' stop segment with an origin of Atlanta, Georgia, and a destination of Tampa, Florida.

B. Return trip

.

- A Type 7 record with an origin of Tampa, Florida, and a destination of Detroit, Michigan (main return record).
- A Type 6 record for the return stop segment with an origin of Tampa, Florida, and a destination of Lexington, Kentucky.
- A Type 6 record for the return stop segment with an origin of Lexington, Kentucky, and a destination of Detroit, Michigan.

If a respondent reported an air trip to Atlanta, Georgia, from his/her home in San Francisco with a 'to' and a return stop in Salt Lake City, the trip would contain the following record types.

A. Going 'to' trip

- A Type 1 trip record with an origin of San Francisco (the respondent's home address) and a destination of Atlanta, Georgia (main 'to' record).
- A Type 2 record with an origin of San Francisco (the respondent's home address) and a destination (departure station) of San Francisco International Airport (SFO).
- A Type 4 record with a 'to' stop segment with an origin of San Francisco International Airport (SFO) and a destination of Salt Lake City Airport (SLC).
- A Type 4 record with a 'to' stop segment with an origin of Salt Lake City Airport (SLC) and a destination of Atlanta, Georgia.

B. Return trip

- A Type 7 record with an origin of Atlanta, Georgia, and a destination of San Francisco.
- A Type 6 record with an origin of Atlanta, Georgia, and a destination of Salt Lake City.
- A Type 6 record with an origin of Salt Lake City and a destination of San Francisco.

2.2 DATA SCREENING AND PRE-PROCESSING STAGE

After data was transferred from the Census to ORNL, several data screening and pre-processing activities were performed. First, initial quick summary statistics were generated and reviewed to gather information about the data set (e.g., number of trips, types of trips, modal shares) and to identify any extreme values or obvious errors for further examination. Next, an automatic checking and editing computer program, AUTOFIX, was utilized to systematically correct mistakes in the trip file. AUTOFIX is a tool developed by ORNL which used several databases built from knowledge learned during the processing of the Wave one - Cycle one data. These databases were updated continuously as new lessons were learned throughout the entire ATS data processing life cycle. After the first few iterations of updates on these knowledge databases, this utility program generally could fix about 70 percent of errors or missing zip codes in the Wave-Cycle data set, systematically.

Depending upon whether specialized checking/editing was needed, trips in each Wave-Cycle data file were separated into subgroups. For example, moving trips and commuter trips were handled differently because of the nature of these trips. Trips with add-stop records required manual checking and editing to insert, or re-sort, intermediate stop locations before trip distance could be calculated. Initial editing was also needed for trips without zip codes (i.e., "uncodable" by the Census) so that geographic locations needed for the calculation of distances could be identified. During the early data processing stage, it was also found that trip segments within the state of Hawaii and trips with waterway reported as their main mode of transportation had higher likelihood of error, which required further checking and editing. These types of trips were pulled out from each data file to form special groups and underwent specific checking and editing.

Commuter Trips:
These trips should have only O-D and no stops. Checks were made for unknown destination city/destination zip; locating proper place name/place zip from various sources, and adding new variable(s) to the trip record. The "commuter mode of transportation" for distance estimation was used; re-coded "other" to major modes of transportation, if needed. Great Circle Distance (GCD) and route-based or estimated distance were provided for these records. If the calculated one-way distance was less than 100 miles, a delete flag was placed on the record.

Mover Trips:
This is the only trip type in ATS that is one way. Checks were made for unknown destination/stop city/zip; locating place name/zip; and adding new variables to the trip record. Foreign country names for the foreign cities were identified, if not already included. Foreign country codes (new variable) were added to all foreign countries. The "mover mode of transportation" was used for distance calculations; the "other mode of transportation" was re-categorized to the four main mode selections (highway, air, train and water). If the mode of transportation for a moving trip was "air", highway was assumed for its associated record Types 11 and 12, unless indicated otherwise. Record Type 10 had the GCD from origin to destination

(direct), the estimated distance from O to D (direct), the sum of segment GCDs from O to D, and the sum of estimated segment distances from O to D. All other record types (11, 12, and 13) had one GCD and one estimated distance for each record. If the estimated total one-way distance was less than 100 miles, a delete flag was placed on the Type 10 record.

Regular Long-Distance Trips:
For Add-stop cases, each individual trip was examined to identify possible/reasonable stop sequence for the trip. New segment records (Types 4 or 6) were added to the trip. The new segment record flag was set to "on" for all newly created segment records and for all those with new stop numbers (due to re-sequencing of stops). The remaining Type 8 records were discarded if the process to "relocate" this information failed. Furthermore, Type 8 records were not included in the final result file sent back to Census.

Checks were made for unknown destination/stop, to locate proper place name/zip, and to add new variables to the trip record. Foreign country names were identified for the foreign cities, if not already included. Foreign country codes (new variable) were added to all foreign countries. Frequencies for all variables were generated and reviewed to try to identify any unusual/unreasonable cases. Trip data was reviewed by grouping trips in one or more of the following ways: (1) with stops vs. non-stops, (2) single mode vs. multiple modes, (3) auto only vs. others, etc. The "other" mode of transportation was re-categorized to the four main mode selections (highway, air, train, and water). If the main mode of transportation for a given trip was "air", "train", or "ship/ferry/boat/etc", highway was assumed for its associated Type 2 record, unless it was reported otherwise. The main mode of transportation was assumed for the entire trip except for cases where (1) a change of mode was indicated, (2) the main mode was unreasonable for a segment, or (3) variables TRNSP019-024 provided other mode information that could be used. Record Type 1 had a GCD from origin to destination (direct), the estimated distance from O to D (direct), the sum of segment GCDs from O to D, and the sum of estimated distances from O to D. All other record types (2, 3, 4, 5, 6, & 7) had one GCD and one estimated distance for each record. If the estimated total one-way distances for both the outbound trip (i.e., origin to destination) and for the return trip (i.e., destination back to origin) were less than 100 miles, a delete flag was placed on record Type 1.

2.3 DISTANCE CALCULATION STAGE

Before the distance calculation could be started, all data files were separated by mode of transportation used for the trips. Trips exclusively by highway (i.e., auto or bus trips) went directly to the highway routing distance program for processing. Initial results from the distance calculation were checked to identify possible errors, particularly in mis-placed geographic locations, reasonableness of the highway mode used, and the order of intermediate stops along the journey. Editing on these segment records was then conducted, and they were resubmitted for distance calculation. This process was repeated as

needed until no apparent errors were found. Detailed description on the highway network database and highway routing algorithm are further discussed in Section 5.

Multi-modal trips which utilized air, train, or boat/ship in part of the reported trips were processed separately. An automatic program was used to perform the computer-aided checking and editing task. This process also allowed a "human-in-the-loop" approach to review trip records when making the editing decisions. Checking on the reasonableness of the given mode used as well as identifying the geographic locations of the reported airport, water port, or train station were the main functions performed under this routine. Missing airports, piers, or train stations in trip segments were assigned, generally, to the nearest location. Further discussion on this process can be found in Section 4.1.2.

After the checking and editing for these multi-modal trip records was completed, all non-highway segments were processed by distance calculation algorithms that used the appropriate mode-specific network databases.   Segments that used highway mode were sent to the highway routing distance program. All distance results associated with each given trip were reassembled before further checking under the post-processing.

To summarize activities under this stage, all trip records, whether in special groups or not, were processed identically as described above. Checking and manual editing, as necessary, were performed before they were sent to the appropriate distance calculation programs.   Distance results were checked for all files, trip segments were manually edited as needed, and distances were re-calculated.

2.4 QUALITY CONTROL AND POST-PROCESSING STAGE

Quality assurance and quality control for edited data and distance results were performed throughout the data processing activity at ORNL. The checking/editing and distance calculation process cycles were repeated as needed for each data file. Circuity factors, defined as the estimated distance divided by the great-circle distance, for all trips were reviewed to identify extreme or unusual cases. These trips were further checked to ensure the accuracy of geographic locations as well as logic in the trip segments. The checking/editing, distance calculation, and post-process checking steps were repeated as needed. Problematic records which could not be resolved within a reasonable time (e.g., due to resource limitations) were flagged for deletion.

Though ATS included trips with final destinations in Mexico, Canada, or overseas, domestic miles traveled was the main focus of the survey. Before files were returned to Census, ORNL generated new records for cases with border crossings or mode changes. The output file to Census was then compiled and converted to a SAS transport formatted data set. The file was then copied to a tape and sent by Federal Express to Census as described above. Sections 3 and 4 describe the checking/editing and QA processes in more detail.

More detailed information on data screening and pre-processing for each sub-group is provided below.

Commuter and Moving Trips:
100 percent sample checking and editing was performed. Checks were made for missing origin and destination and for missing or reasonableness of the mode. Edits/imputations to records were made on hardcopy printout, then on a computer file. Records were re-checked and re-edited. The file was prepared for distance calculation. Distance was calculated for each trip, and QC checks were performed on the results. If problems were identified, records were reviewed and edited, then distance was re-estimated.

The screening and pre-processing for moving trips was similar to that of commuter trips with additional checking and editing on stop sequence, when needed.

Add-Stop Trips:
The travel route was plotted with mapping software (in most cases *Map&Go*) to identify a reasonable location for the "Add Stop." The given number of nights were added to the appropriate records. The mode of transportation reported for the segment was checked, and stop segments were re-sequenced when necessary. Records were checked and edited, re-checked and re-edited. The file was prepared for distance calculation. Distance was calculated for each trip, and QC checks were performed on the results. If problems were identified, records were reviewed and edited, then estimated distance was re-estimated.

Uncodable (Zips with '99999', 'D','R','DK', or blank) Cases:
100 percent sample checking and editing was performed. The procedure for imputing zip codes was followed (see Section 3). The route was checked with mapping software; place names were checked with atlases, zip code databases, phone directories, the Internet, and the Geographic Names Information System (GNIS). The segment sequence was checked, edited, and reviewed. The file was prepared for distance calculation. Distance was calculated for each trip, and QC checks were performed on the results. If problems were identified, records were reviewed and edited, then distance was re-estimated.

Waterway as the Reported Main Mode of Transportation:
100 percent sample checking and editing was performed. The reasonableness of routes for the given mode was checked. The mode or missing location was edited or imputed, as needed. Stop segments were re-sequenced, when necessary. The file was prepared for distance calculation. Distance was calculated for each trip, and QC checks were performed on the results. If problems were identified, records were reviewed and edited, then distance was re-estimated.

Hawaii to Hawaii (Alaska to Alaska, Alaska to U.S., U.S. to Alaska) were processed similar to waterway trips.

Air Trips:
Records were checked for airport locations. The reasonableness of the mode for a given segment was checked. The file was prepared for distance calculation. Distance was calculated for each trip, and QC checks were performed on the results. If problems were identified, records were reviewed and edited, then distance was re-estimated.

Non-Stops and Others:
The file was prepared for distance calculation. Distance was calculated for each trip, and QC checks were performed on the results. If problems were identified, records were reviewed and edited, then distance was re-estimated.

3. DETAILED PROCEDURES FOR DATA SCREENING AND PRE-PROCESSING

This section provides detailed descriptions of the procedures developed by ORNL for data processing activities under the ATS project. Data editing and imputation of missing geographic locations were required in order to estimate travel distance for trips reported in the survey. Some of these procedures apply in general to all records, regardless of the record type. Others are specific for particular types of records (e.g, records about a household move or records for the "access" segment).

3.1 GENERAL PROCEDURES

The procedures discussed in this section applied to all records in the ATS trip file.

- Trip records were identified by a group of fields. These fields were included with segment records <u>at all times</u> so that given trips could be properly identified. The following fields were unique identifiers for ATS segment records.

    CaseID (identified the household)
    Persons (identified the person in a given household)
    Mdest (i.e., trip number, which identified a given trip taken by the given person)
    Rectype (identified the type of trips or segments and its direction, whether going or return, when appropriate)
    Stopnum (identified the stop segment number for a given trip)

- ORNL was always "conservative" with any changes that were made to data files. Changes were made only to those data that were obviously wrong, unreasonable, or out-of-sequence. Original survey responses were always preserved.

- Due to the lack of available network databases for subway/commuter and rail/elevated rail, segments with these types of transportation (usually for the "access" segment, i.e., Type 2 record) were routed on the highway network for distance estimation.

- As an initial examination of the trip data, frequency tables (or other descriptive statistics) of variables in the trip segment file were reviewed carefully to detect any unusual or unreasonable cases.

- New variables were added to the distance output file to transfer information found from the data checking and editing process back to the Census. Data flags were also added to the file to signal changes made to a record (see Appendices A and B).

- Hawaii and other domestic locations were coded as 'FC' in the state name field in some trip records. Consequently, zip codes for these given locations were unspecified. These cases usually specified 'Hawaii,' etc., as their foreign country name so they were relatively easy to spot and fix.

- Puerto Rico, Guam, Samoa, and other U.S. territories were treated as foreign countries in the ATS project. Foreign country codes were required for these locations. Only GCD was calculated for foreign segments (i.e., segments that fell entirely outside the United States).

- When air was reported as the main mode, the following procedure was followed. The trip destination was not reordered when the respondent reported the arrival airport city as his/her destination and the actual place he/she visited as a side trip for the trip. From the transportation point of view this was a correct response, because the main mode was air and the air trip ended at the arrival airport.

- For all domestic trip segment records, both GCD and estimated distance (algorithm distance) was provided. Highway segments were routed by algorithm on the updated highway network with the auto impedance function. The distance for air trip segments and water trip segments was estimated from port to port based on GCD. Amtrak trip segments were estimated based on available databases on Amtrak stations and Amtrak routes.

- The zip code field was examined for non-numeric characters other than the 'D', 'DK', and 'R'. Census coders used 'D' for 'Don't Know' and 'R' for 'Refused'. These were 'uncodable' cases, not errors.

- After the distance estimation process was completed, all segments for a given trip were reviewed together. The GCD for the "direct" O-D (i.e., non-stop distance from home to the final destination) was compared to the sum of segment distances for the same O-D pair. If a discrepancy in the distances appeared large (compared to a certain "pre-determined" ratio for data screening purposes), a closer check of the entire trip was made.

- ORNL marked the trip for deletion if both of its total estimated one-way O-D and D-O distances were less than 100 miles.

- When a county name was reported instead of the actual city name, the most populated zip code in the given county was used as the county's zip code. An alternate method of assigning a randomly selected zip code within the given county was also considered. This approach seemed statistically less biased. The drawback of this method, however, was that it could not easily reproduce the same zip code for future trips made by the same person/household for the same O-D pair. The additional effort required for this alternate method offset the gain of using it.

- All trips with a foreign country origin were out-of-scope trips and were marked for deletion.

3.2 RECODING AND EDITING OF THE MAIN MODE

The following procedures were used when a change of segment-mode or trip-mode was required.

- If the mode of transportation (e.g., the value in the Census's field TRNSPRT1) was coded as "other" and additional information was available in the separate description variable (e.g., the field TRNSPR1S), the procedure provided in Tables 3–5 was followed. Table 3 shows the procedures for recoding values when the field TOSTATNS was "other" and additional information was available in the description variable. Table 4 shows the procedures for recoding values for the field FRSTATNS and Table 5 for the field TRNSPRT1S. These lists were compiled based on data given by respondents during the Wave-one Cycle-one interviews.

- If the field TRNSPRT2 (main mode used for the return trip) was unspecified, either blank or coded as '18' for others, the same code as specified in TRNSPRT1 (main mode used for going trip) was used. That is, it was assumed that returning trips used the same type of mode of transportation as the departing trips.

- For trips where the main mode of transportation was <u>cruise ship</u>, some editing and re-coding of variables as well as re-ordering of stop-sequence was necessary. Generally, these type of records were coded with "other" as their mode of transportation in the station (variable TOSTATN) and then described in the TOSTATNS field as "commercial air" or "airplane." The reported "station", in fact, should have been a "stop" where the change of mode to cruise ship occurred. These travelers typically made an air trip from home to the reported "station" to get on the cruise ship. Under these circumstances, the correct "station" was changed to the airport where the respondent departed from to begin his/her trip. The TOSTATN field value, typically, should be "auto", and the FRSTATN field value could be taxi, shuttle bus, or an equivalent. The originally reported "station", thus, was changed to a "stop" with a mode change from air to cruise ship at this location. Consequently, stop numbers for all other reported stops in this trip were re-ordered to reflect these changes.

- If the mode of transportation, TRNSPRT1 field, for record Type 1 was missing or coded as "other" without specification, it was assumed to be highway for all land trips and to be air for all others, unless other information indicated differently (e.g., apparent cruise ship trips).

- All trip records with at least one end of its segment located in "HI", "AK", "PR", or "FC" (i.e., foreign country) were reviewed carefully to determine the reasonableness of the given mode of transportation or to decide the missing mode of transportation for the segment in question. For example, special attention was paid to travel made between Hawaii islands, between Hawaii and the U.S. mainland, and between Alaska and the U.S. mainland. The *main* mode of transportation used for these trips should not be highway related unless the trip was between Alaska and the United States. In that case, the trip data would need to be reviewed.

Table 3. Recoding "Other Specified" responses for the variable TOSTATNS
(transportation used from home to the departure station/airport/port/terminal)

Recode to Value	Value Description	Description Given by Respondent
1	A car, truck, or van parked at the airport/station/port/terminal	Company car parked at airport Friend's car, truck, or van parked at airport Relative's car, truck, or van parked at airport Own car parked at office, which is close to airport Rental car COE worker City vehicle
2	Car, truck, or van dropped off by another person	Co-worker dropped respondent off Friend dropped respondent off Relative dropped respondent off Ambulance Mini bus privately owned Car parked at hotel Car Service RV
3	Motorcycle, moped, or motor bicycle	
4	Taxi	Chartered car Hired driver Took car to friend's house and taxi to airport Car service
5	Limousine or shuttle bus	Airport bus Airport shuttle Airport van Charter bus Charter trip Coach bus Courtesy van Hotel van Military bus Military transport vehicle Regular bus service like public bus but not School bus Shuttle van Tour bus Travel company's bus
6	City-to-city bus	
7	Subway/elevated rail/commuter rail	Train MBTA train
8	Walked	

Table 3. (Continued)

Recode to Value	Value Description	Description Given by Respondent
9	Other type of transportation	4 wheeler ATV
		Air taxi
		Airplane
		Bicycle
		Boat
		Commercial plane
		Corporate plane
		Dog sled
		Ferry
		Motor launch
		Snowmobile
		Snowmachine

Table 4. Recoding "Other Specified" responses for the variable FRSTATNS (transportation used from arrival station/airport/port/terminal to the final destination)

Recode to Value	Value Description	Description Given by Respondents
1	Picked up in a privately owned car, truck, or van	Ambulance
		Company car
		Picked up by relative
		Picked up by friend
		picked up by coworker
2	Rented car, truck, or van	
3	Taxi	Chartered car
		hired driver
		Official government taxi service
4	Limousine or shuttle bus	Airport bus
		Airport shuttle
		Airport van
		Chartered bus
		Chartered trip
		Courtesy van
		Hotel van
		Military bus
		Military transport vehicle
		School bus
		Tour bus
5	Public bus	City-to-city bus
6	Subway/elevated rail/commuter rail	Train
7	Walked	

Table 4. (Continued)

Recode to Value	Value Description	Description Given by Respondent
8	Other type of transportation	Bicycle
		Boat
		Commercial plane
		Corporate plane
		Dog sled
		Ferry
		Motor launch
		Snowmobile
		Snowmachine

Table 5. Recoding "Other Specified" responses for all other mode variables (TRNSPRT1, etc.)

Recode to Value	Value Description	Description Given by Respondent
1	Car, pickup truck, or van	Ambulance
		Blazer
		Church car/van
		Company car/van
		Four wheel drive vehicle
		Friend's car
		Funeral coach
		Government car
		Government-issued auto
		Hearse
		Jeep
		Limousine
		Parent's car
		Roll back
		Utility vehicles
2	Other truck	Army truck
		Company dump truck
		Company truck
		Flatbed truck
		Logging truck
		Military truck
		Semi truck
		Tractor trailer
		U-haul
		18 wheeler
3	Rental car, truck, or van	Charter van
4	Commercial airplane	Air
		Airplane
		Charter plane

Table 5. (Continued)

Recode to Value	Value Description	Description Given by Respondent
5	Corporate/personal airplane	Government plane
		Helicopter
		Military airplane
		Medivac helicopter
		Naval airplane
		Private plane
6	City-to-city bus	Greyhound
		Commercial bus line
7	Charter bus or tour bus	Church bus
		Military bus/shuttle
		Privately owned bus
		Shuttle bus
8	School bus	Boy scout bus
		Girl scout bus
		School Van
9	Train	Amtrak
		Bart
		El
		Metro
		Subway
		T
10	Taxi	
11	Ship or boat	Cargo ship
12	Cruise ship	
13	Passenger line or ferry	Water taxi
14	Recreational boat, sailboat, pleasure boat, or yacht	
15	Recreational vehicle or motor home	Motor home
		Mobile unit
16	Bicycle	
17	Motorcycle, moped, or motor bicycle	Solar-powered go-cart
18	Other	Dogsled
		Snowmobile

3.3 PROCEDURES FOR UNKNOWN OR MISSING GEOGRAPHIC LOCATIONS

The following procedures were followed for situations in which the origin, destination, or stop locations are unknown. In most cases, these procedures were followed in the order they are listed.

3.3.1 Unknown Origin

The following procedures apply to trips, other than moving trips, where part or all of the origin location (state, city, or zip code) was unidentified. The origin of these trips, typically, is the sample address. Therefore it was less likely to be entirely missing. Most of the errors occurred because of typos in the zip code or 2-digit state name.

1. Other trips reported by the same person or someone from the same household were checked to find needed information.
2. If given, state and city names were checked for possible misspelled names. For example, checks were made for the same place name in different states or sound-alike names to locate the missing zip code (e.g., mistakes such as "MI" for Mississippi, "MS" for Missouri, "AL" for Alaska, or "AK" for Arkansas, were frequently found).
3. ORNL provided the Census a list of records with unknown origins to check with other sources at the Census when possible.
4. As a last resort, if the state was specified and the city was unknown, the zip code provided in Table 6 for the given state was used. For trips that were made from the old address (prior to the moving trip), the same procedures described in Section 3.3.2 were applied. Note that the zip code given under column heading "Primary Zip" in Table 6 was used when imputing a zip code for a given state. The "alternate zip" was selected only if the primary city happened to be the same city as the other end of the trip segment in question.

Table 6. Zip Code Used for Imputation (Given State)

State	Primary City	Primary Zip	Alternate City	Alternate Zip
AK	ANCHORAGE	99504	JUNEAU	99801
AL	BIRMINGHAM	35215	MONTGOMERY	36116
AR	LITTLE ROCK	72209	PINE BLUFF	71603
AZ	PHOENIX	85023	TUCSON	85706
CA	LOS ANGELES	90011	SAN FRANCISCO	94110
CO	DENVER	80123	PUEBLO	81001
CT	HARTFORD	06106	NEW HAVEN	06511
DC	WASHINGTON	20011		
DE	NEWARK	19711	DOVER	19904
FL	ORLANDO	32808	MIAMI	33012
GA	ATLANTA	30318	MACON	31206
HI	HONOLULU	96818	HILO	96720
IA	DES MOINES	50317	CEDAR RAPIDS	52402

Table 6. (Continued)

State	Primary City	Primary Zip	Alternate City	Alternate Zip
IL	CHICAGO	60623	SPRINGFIELD	62702
IN	INDIANAPOLIS	46227	SOUTH BEND	46614
KS	MANHATTAN	66502	WICHITA	67212
KY	LOUISVILLE	40214	LEXINGTON	40517
LA	NEW ORLEANS	70117	LAFAYETTE	70506
MA	BOSTON	02154	SPRINGFIELD	01109
MD	BALTIMORE	21215	HAGERSTOWN	21740
ME	PORTLAND	04103	BANGOR	04401
MI	DETROIT	48227	LANSING	48911
MN	MINNEAPOLIS	55423	ST CLOUD	56396
MO	SAINT LOUIS	63136	COLUMBIA	65203
MS	JACKSON	39209	HATTIESBURG	39401
MT	HELENA	59601	BILLINGS	59102
NC	GREENSBORO	27406	CHARLOTTE	28205
ND	BISMARCK	58501	FARGO	58103
NE	OMAHA	68104	LINCOLN	68502
NH	MANCHESTER	03103	LEBANON	03766
NJ	TRENTON	08618	NEWARK	07104
NM	ALBUQUERQUE	87105	LAS CRUCES	88005
NV	LAS VEGAS	89121	RENO	89502
NY	NEW YORK	10025	ALBANY	12203
OH	COLUMBUS	43229	CLEVELAND	44107
OK	OKLAHOMA CITY	73160	TULSA	74133
OR	PORTLAND	97206	EUGENE	97405
PA	PHILADELPHIA	19143	PITTSBURGH	15235
RI	PROVIDENCE	02908	NEWPORT	02840
SC	COLUMBIA	29203	CHARLESTON	29407
SD	RAPID CITY	57701	SIOUX FALLS	57103
TN	NASHVILLE	37211	KNOXVILLE	37920
TX	DALLAS	75217	HOUSTON	77036
UT	SALT LAKE CITY	84118	SALINA	84654
VA	RICHMOND	23223	LYNCHBURG	24502
VT	BURLINGTON	05401	RUTLAND	05701
WA	SEATTLE	98115	YAKIMA	98902

Table 6. (Continued)

State	Primary City	Primary Zip	Alternate City	Alternate Zip
WI	MILWAUKEE	53209	MADISON	53711
WV	CHARLESTON	25302	MORGONTOWN	26505
WY	CASPER	82604	CHEYENNE	82001

3.3.2 Unknown Domestic Destination

The following procedures applied when all or part of the destination location was unknown, but the location was inside the United States.

- Cases with city and state names but no zip code was specified

  1. State and city names were checked for possible misspelled names, the same place name in different states, or sound-alike names to locate the missing zip code.
  2. Geographic place name database, map software, or the telephone directory was used to locate destinations reported with parks, resorts, business names, etc.
  3. If a place could be found with its x-y coordinates but no zip code was available, then
     (a.) the zip code that the x-y coordinates were located in, or
     (b.) the nearest city's zip code
     was used.
  4. If all of the above failed and no such city could be found in the given state, we assumed the city name was reported correctly and followed the procedures described in the "state unknown" case to determine a state name.
  5. For cases with unspecific destinations, such as "Pump Station 12", "Oil Field", or "Oil Rim in the Gulf", etc., we used an "educated guess" to find a proper place for the purpose of distance calculations. Unlike the general rules for unknown cities, these destinations were less likely to be in a large city.

- Cases with state name but no specific city name

  1. Available information (e.g., other trips made by the person or others in the household) was checked to identify a city location for the trip, if possible.
  2. The largest/most-popular city selected from the pre-compiled list in Table 6 was used for the given state.

- Cases with city name but no specific state name

1. Available information (e.g., other trips made by the person or others in the household) was checked to identify a state location for the trip, if possible
2. States with the given city name were identified and the most likely state for the trip was determined.
3. If more than one state was possible for the given trip, the state with the larger (or largest) city was used for the state name.

- Cases with unknown city and unknown state

1. Other trip records in the same household were reviewed to make an "educated guess" on destination location;
2. The state that was most commonly visited by all persons traveling from the same area (same zip code or same town) in the given Wave and Cycle combination (e.g., W1C1) was used. The resulting state was used to find a zip code from Table 6 and entered in the trip record as an imputed destination zip.

3.3.3 Unknown Domestic Stop

The following procedures were used when all or part of a stop location was unknown. Travel route was always reviewed with map software when identifying the missing stop.

- Cases with city and state names but no zip code was specified

1. State and city names were checked for possible misspelled names, the same place name in different states, or sound-alike names to locate the missing zip code.
2. Geographic place name database, map software, or the telephone directory was used to locate destinations reported with parks, resorts, business names, etc.
3. If a place could be found with its x-y coordinates but no zip code was available, then
   (a.) the zip code that the x-y coordinates were located in was used or
   (b.) the nearest city's zip code.
4. If the given city did not exist in the given state, we assumed the city name was reported correctly and followed the procedures described in the "state unknown" case to determine a state name.

- Cases with state name but no specific city name

1. Available information from the trip in question or other trips made by the household was used to identify a possible stop location for the trip (e.g., similar trips made by the same person or the same trip reported by another person in the same household).

2. The route was checked to locate possible stops in the given state; the larger cities <u>along the route</u> were selected. The largest/most-popular city selected from the pre-compiled list in Table 6 was used for the given state, if the selected stop was reasonably located.
3. If all attempts to locate the city failed, the stop was marked for deletion.

- Cases with city name but no specific state name

1. States with the given city name were identified and the most likely state was determined (i.e., on the travel route) for the trip;
2. If more than one state was possible for the given trip, the state with the larger (or largest) city was used for state name.
3. If no reasonable location could be identified, the stop was marked for deletion and stop numbers were resequenced, if necessary.

- Cases with unknown city and unknown state

1. Other trip data reported by the same household was reviewed to make an "educated guess" on a possible stop location.
2. If this attempt failed, the stop record was marked for deletion, and the stop sequence was reordered if necessary.

- Cases where destination or stops reported using county name instead of city name

1. The county name in the given state was located and the place name with the most populated zip code within the county was selected (should be on or reasonably near the travel route).
2. If the county name did not exist in the given state or if the state name was missing, all states that the given county name could be found in were identified; the most likely county, city, zip for the trip was selected (should be on or reasonably near the travel route).
3. Possible misspelled county name or sound-alike names were also checked.
4. If all these attempts failed, the stop record was marked for deletion.

3.3.4 Unknown Foreign Destination/Stop

- Cases with known city and country names

1. The foreign city's x-y coordinates from the Digital Chart of the World database were used instead of a zip code.
2. The 3-digit foreign country code from the Standard Foreign Country Code database was added to the output file.
3. The foreign country code and a 4-digit sequential number were used to form an ID code for the given foreign city for distance calculation.

- Cases with known city but unknown country names

    1. Information in the other trip records from the same household were reviewed to determine possible country (countries) for the given city.
    2. The most popular/largest country with a city name as given was selected if multiple choices were found. The 3-digit foreign country code and the x-y coordinates for the city were used for distance estimation.
    3. If no reasonable choice for a stop could be identified, the stop record was marked for deletion, and stop segments were resequenced, if necessary.

- Cases with known country name but unknown city

    1. The trip record was reviewed to determine possible city location for the given country.
    2. The major/largest city in that country was used to get x-y coordinates for distance calculation.

- Cases with unknown city and country names

    1. If the unknown location was a destination, all information in the given trip or other trips taken by the same household was reviewed to determine a possible location.
    2. If the destination location still could not be determined, the most frequently visited country by persons traveling from the same origin city/state in the current ATS Wave-Cycle file was used.
    3. If the unknown location was a stop, the stop record was marked for deletion, and stop segments were resequenced, if necessary.

3.4 PROCEDURES APPLIED TO SPECIFIC RECORD TYPES

.
3.4.1 Commuter Trips (Record Type 9)

A record marked as a Type 9 indicated that it was a reported commuter trip. Information on stops were not collected for this type of trip. Trip data generally included city name, state, and zip code for both origin (home) and destination, foreign country name if appropriate, and the main mode of transportation used for the commute. The following procedures applied to commuting trip records. Generally, if reported, there was only one commuter trip for each person in a given household.

- Cases with unknown mode

    When the mode of transportation for a commuter trip was not reported, highway was assumed unless other information indicated differently (e.g., travel between Hawaii islands).

- Cases with unknown origins

    1. The origin of other trips made by the same person or member of the same household was used to replace the missing origin.

    2. The Case IDs for these types of trips were provided to Census so that further information, if it existed, could be extracted from other files at the Census.

- Cases with unknown destinations

    1. The missing destination was replaced with the destination given by another commuter in the same household.
    2. The missing state was replaced with the state given by other commuters living in the same city (especially for those who live in Alaska), and the zip was imputed using Table 6.
    3. The missing destination was replaced with the nearest large city located outside a 75-mile radius of the origin.

If both the origin and destination of a given commuter trip were missing, the trip was marked for deletion.

3.4.2 Moving Trips (Record Types 10, 11, 12, and 13)

Moving trips are those where the person was moved from another location into the sample address. Commuter trips and moving trips were the only trips in ATS files that were one-way.

- Cases with unknown mode

    The mode of transportation used for moving Type 11 records was assumed as highway unless information gathered from the trip record indicated otherwise.

- Cases with unknown destination

    The origin from non-moving trips made by members in a household after the move date was used to replace unknown destinations occurring in moving trips made by persons from the same household. Note that date information was not available in the trip file for distance calculation. A list of Case IDs were compiled so that other records for the household could be checked at the Census, if necessary.

- Cases with unknown origins

1. All information in the given trip or other trips taken by the same household were reviewed to determine possible origin location.
2. As the last resort, we used the given destination state to randomly select one of the top five out-moving states from Table 7.[4] For example, if the origin of a moving trip was unknown and the destination trip state was Alabama, we would randomly select a state from the first row (AL) in Table 7 (i.e., FL, GA, TX, TN, or MS), but only if all other attempts to find the origin had failed. Once the state was determined, the zip code from Table 6 for the selected state was used.

Table 7. State of Residence in 1990 by Top 5 Out-Moving
State of Residence in 1985

State 1990	Top Out-moving State 1985	2nd Out-moving State 1985	3rd Out-moving State 1985	4th Out-moving State 1985	5th Out-moving State 1985
AL	FL	GA	TX	TN	MS
AK	WA	CA	TX	OR	CO
AZ	CA	TX	IL	CO	NY
AR	TX	OK	CA	MO	LA
CA	NY	IL	AZ	WA	CO
CO	CA	TX	IL	AZ	NE
CT	NY	MA	NJ	CA	FL
DE	PA	MD	NJ	NY	TX
DC	MD	VA	NY	CA	PA
FL	NY	NJ	OH	TX	IL
GA	FL	TX	AL	TN	NC
HI	CA	TX	WA	VA	FL
ID	CA	WA	OR	UT	MT
IL	FL	CA	TX	WI	IN
IN	IL	OH	MI	KY	TX
IA	IL	NE	MN	TX	CA
KS	MO	TX	OK	CA	CO
KY	OH	IN	FL	TN	TX
LA	TX	MS	CA	FL	AL

[4]Table 7 was compiled based on information extracted from the *1990 Selected Place of Birth Statistics for States*, CPH-L-121, Kirstin A. Hansen, Journey to Work and Migration Statistics Branch, Population Division, U.S. Bureau of the Census, Washington, DC 20233.

Table 7. Continued

State 1990	Top Out-moving State 1985	2nd Out-moving State 1985	3rd Out-moving State 1985	4th Out-moving State 1985	5th Out-moving State 1985
MA	NY	CT	CA	NH	NJ
MI	OH	IL	TX	FL	CA
MN	WI	ND	IL	IA	CA
MS	LA	TN	TX	AL	IL
MO	IL	KS	TX	CA	OK
MT	CA	WA	WY	CO	ID
NE	IA	CO	CA	TX	KS
NV	CA	AZ	UT	TX	IL
NH	MA	NY	CT	VA	CA
NJ	NY	PA	FL	CA	TX
NM	TX	CA	CO	AZ	OK
NY	NJ	FL	CA	PA	MA
NC	VA	FL	NY	SC	TX
ND	MN	MT	SD	CA	TX
OH	FL	MI	PA	TX	CA
OK	TX	CA	KS	AR	MO
OR	CA	WA	ID	TX	AK
PA	NJ	NY	MD	OH	FL
RI	MA	NY	CT	CA	NJ
SC	NC	GA	FL	NY	VA
SD	MN	IA	ND	NE	CA
TN	TX	FL	GA	KY	IL
TX	CA	LA	OK	IL	FL
UT	CA	ID	AZ	CO	TX
VT	MA	NY	NH	CT	NJ
VA	MD	NY	PA	CA	NC
WA	CA	ID	OR	TX	AK
WV	OH	VA	MD	PA	FL
WI	IL	MN	CA	MI	TX
WY	CO	MT	CA	UT	TX

Source: 1990 Selected Place of Birth Statistics for States, CPH-L-121, Kristin A. Hansen, Journey to Work and Migration Statistics Branch, Population Division, U.S. Bureau of the Census, Washington, DC 20233.

Note that a moving trip can occur before or after other trips reported by the person or others in the same household. Furthermore, records that fail to include trip origins generally were moving trips or trips made from the respondent's old address. These trips are not 'repeatable' (e.g., the same person

can't be moved from the old address again). Therefore, the ability to reproduce the same imputed location for an unknown origin in future trips made by the same person was not a concern.

3.4.3 Regular Long-Distance Trips (Record Types 1-8)

These trips are those that are not reported as commuting or moving trips. These trips may or may not include stops along the way. For definition of these record types, please refer to Section 2.1.2.

3.4.3.1 Side Trips (Type 5)

By definition, a side trip started from the destination and came back to the same destination after the survey respondent visited another place(s). Several side trips could occur during the stay at a destination. However, there was no indication as to how the survey respondent actually traveled (i.e., returned to the destination after one side trip or traveled from one place to the other before returning to the destination). The distance for a Type 5 record was calculated for the O-D as given in the segment file. That is, no attempt was made to review or readjust the logic of side trip segments. Type 5 distance was not included in the total trip distance calculation per BTS requirement.

- Only side trips that contained obvious errors in logic were corrected. The following is an example of such a case.

  A respondent reported an auto trip to Hartford, CT with side trips to San Juan, St. Croix, St. Thomas, and San Juan.   From reviewing the information in variables TRNSP019-TRNSP24 in the Type 1 record, the modes used (for more than 75 miles) were listed as "01" (auto), "04" (air), and "12" (cruise ship). The logic of the reported trip is obviously wrong. We changed Hartford to a stop, where the traveler changed the mode from auto to air, and changed San Juan to the final destination where a mode change from air to cruise ship occurred. The side trips were made from San Juan.

3.4.3.2 Border Crossing

The following procedures applied to trips where the respondent crossed the U.S. border into a foreign country.

- When a border crossing occurred in routing a highway segment, the segment was broken into two records at the border crossing place so that domestic and international distances could be separately accounted for. This process was systematically conducted by the highway distance estimation procedure. In the case of air travel, the separation of domestic and international segments was made at the exit (i.e., the last U.S. airport on the way to a foreign country) or the re-entry (i.e., the first U.S. airport on the way back from the foreign country) airports.

- The distance calculation algorithm reported the domestic segment GCD and its estimated distance, international segment GCD and its estimated distance, and location of the border crossing (city and state, if available). Note that an auto trip between one of the lower 48 states and Alaska would have at least 2 border crossings each way.

3.4.3.3 "Implicit" Mode Changes

The following procedures and explanations applied for mode changes that were not specified by the respondent during the reporting of a trip.

- It is assumed that the respondent used the same airport/station to arrive at the destination and to return home. It is also assumed that the respondent traveled back to the same airport/station that was used as the departure airport/station in his/her going trip. In order to have air/water/train and highway distances properly separated, when a mode change was not reported but required to travel on a given segment, such segment record was split into two records. Each of the two segment records was identified with the same ID with several additional variables. This process was systematically performed under the distance estimation operation. Appendix B in this document provides the definition of these variables.

3.4.3.4 Add Stops (Type 8)

As indicated in Section 2.1.2, Type 8 records needed to be manually examined and added back to the trip record in a logical order. In most cases, stop segments also needed to be re-sequenced accordingly. Rule of thumb and common sense were needed in order to place these stops as accurately as possible. Many of the rules specified in the "General Procedures" section for handling missing information are applicable to the "Add Stops" situations as well. Additional rules and procedures that were used in interpreting "Add Stops" are listed below.

- When an "Add Stop" happened to be at one of the locations that the respondent had already reported for this given trip (destination, stop, or side trip), the reported "Add Stop Nights" was added to its associated variable and no new segment needed to be added.

- If the reported "Add Stop" was listed as "slept in car" or a similar type of response, no new segment was added. The additional night was added to the ROADMISS variable, which represented the number of nights that the traveler spent in transit.

- *Map&Go* software was used to show a possible travel route for the given origin-destination pair, with all known stops in the order as reported. The route was reviewed in order to place the added stop along the selected route. This procedure usually worked well for auto trips.

- If the "Add Stop" was close to the reported destination but not directly on its route, this segment was added as either a stop or a side trip. By Census definition, a side trip starts from the destination and returns to that same city.

- *Map&Go* was used to identify the zip code for the given "add-stop" city as well. When multiple zip codes were available, the 1995 GDT Zip Code file was used to select the most populated zip code for that location.

- Some "Add Stops" were not directly located on the selected route (usually quickest or shortest between two given points); however, it was still added to the record as a stop. This situation occurred when the traveler had a special reason, other than spending the night (e.g., visiting a friend or relative), to stop at a location. This case was especially true for auto trips.

- Adding a stop for an air trip was not always straightforward. The stop might require a change of mode to auto at some airport location to drive to the place indicated by the added stop or it might require a change of mode to cruise ship at a certain port in order to get to the added stop. The logic of the trip was carefully reviewed and followed.

- If the given "Add Stop" was at a city close to the airport/station from which the respondent flew out to begin his/her trip, this add-stop location was inserted to the trip as a stop between the arrival airport and home (i.e., Type 6 segment). This insertion was made to avoid the complication of coding a stop for the Type 2 record, which represented the "access link" from home to the airport/station.

- When several added stops for an <u>auto trip</u> were equally likely to occur in either of the "going" or the "returning" trip directions, the following procedures were used.

    1. If travel in the trip segment in question could be completed within a reasonable time frame, these added stops were placed in the "going" direction.
    2. If the travel could not be completed in a reasonable time frame, then the added stops were split into two groups; one group (in a logical order) was added to each of the travel directions.
    3. When two add-stop cities were relatively close together, especially if the reason for the stop was to "spend the night" (code 12), they were added to both legs of the trip, one for each direction.

- If an "Add Stop" was given without city and state names and if no other information from the trip record indicated otherwise, the additional nights reported for this add-stop location were added to its final destination. No new stops records were added. This was mostly applicable for the trips that used air as the main mode of transportation. For auto trips (or trips with auto as the main mode), if a segment of the given trip appeared to be too long for non-stop travel, the "Add Stop" was imputed

by using *Map&Go* to locate a reasonable location (e.g., larger city) along the route for the segment in question. The additional nights reported for this "Add Stop" were then assigned to the location found.

Note that in an effort to reduce the data processing time for ATS, the "add-stop" process was terminated after the processing of Cycle one data files. This was a trade-off action between the data accuracy and the project resources called for by the BTS.

4. MODIFIED PROCEDURES FOR THE SPEEDY PROCESSING

An in-progress review meeting on ORNL's ATS data processing operation was conducted near the completion of the first Cycle data files (included 3 Waves). In addition to members of the ORNL project team, this meeting was attended by representatives from the BTS and the Census. Although the overall data processing effort produced impressive results with accuracy, it was clear that ORNL's undertaking was tremendously time consuming. In an effort to reduce the data processing time and accelerate the ATS distance calculation operation, the BTS program manager instructed ORNL to modify the original processing procedures used for Cycle one data files so that preliminary ATS results could be produced by an earlier schedule. The modified operation is referred to as "speedy processing" in this report and is discussed in this section.

4.1 CHANGES IN THE NEW PROCESSING METHOD

Major changes under the modified speed-up process include the following.

A.   Eliminated all "Add-stop" records.

Type 8 records were excluded from the ATS trip files before further data processing was done. That is, no attempt was made after Cycle one data processing to insert additional intermediate stops back to the reported trip. This removed most of the location placements and segment re-sequencing needs from the data processing.

B.   Minimized manual checking and editing activities during the data screening and pre-processing operations.

Generally, no investigation or research on missing geographic locations were to be conducted. Manual editing or imputation was to be performed only if the matter of fixing such an error, or a missing value, was obviously simple. Unfixed segment records were marked with an 'X' to reflect that such records were to be skipped from the distance calculation. That is, no distance estimates would be provided for these segments.

C.   Limited checking and re-editing of trip records after the distance calculation processing stage.

Some trip segments might not show apparent mistakes from the initial screening. After distances were computed, unreasonable results associated with inadequacy of the geographic locations or logical problems in segment sequencing might be revealed during the post-processing check. Similar rules on the determination of fixing or skipping such records as those set for the pre-processing were applied. Most of these segments would be "X" marked.

4.2 GENERAL PROCESSING PROCEDURES

The overall procedures for checking, editing, and imputation under the speedy processing were generally similar to those reported in Section 3 with exceptions described in the above section. The level of effort in determining whether to fix, or in attempting to find the fix, however, was dropped considerably. Because the manual checking and editing demand had been reduced, trip files from the Wave-Cycle data sets were not split into as many subgroups as they were previously during the pre-processing stage. Essentially, commuter and moving trips were the only types of records being pulled out from the rest of the trip file and processed separately.

4.2.1 Common Procedures

The following procedures were used in the speedy processing for all ATS data.

- The "autofix" utility program and databases were run on the data files in an attempt to automatically correct as many uncodables as possible.

- Except in a few cases where time permitted and the files were small (e.g., commuters and mover files), uncodables were not pulled out for further review before trip records were processed by the air or highway mode-specific programs.

4.2.2 Mode-Specific Procedures

A. Highway Processing

Records returned from the highway QC program checks and flagged for problems (see Section 5) received a quick review. Under this process, ORNL still attempted to save as many trips and associated stop records as possible.

- Cases with unknown mode

    When the mode of transportation for a given trip was not reported, highway was assumed unless other information indicated differently (e.g., travel between Hawaii islands). When the mode of transportation for a stop was unreported, the main mode of transportation was assumed for this segment unless other information indicated differently.

- Cases with unknown destination

    (1) The following procedures were used for unknown trip destinations in Type 1 records. If the state was reported but not the city, the largest/most-populated zip code selected from the pre-compiled list in Table 6 was used. If the city was reported but not the

state and it was obvious from a quick review of the trip what state that city was located in, the record was edited. If the city/state could not be identified easily, all trip records associated with the given trip were marked with an "X" in the delete flag field, DLTFLAG (see Appendix B). If neither city nor state were reported, all records for that trip were marked with an "X" in the DLTFLAG field. Records that contained an "X" in the DLTFLAG field were skipped during the distance calculations.

(2) <u>If the destination in a stop segment was not reported, the following procedures were used.</u> If the destination could be identified easily by the reviewer (e.g., there was a typo in the city name), this stop record would be edited with a new location. If the destination could not be easily identified by the reviewer, the stop record and all associated stop records of the same type were marked with an "X" in the DLTFLAG field. For example, if the stop was made going to the destination (Type 4 record), all Type 4 records would have an "X" in the DLTFLAG field. Distance would be computed only for the distance from origin to destination for the "going" trip. Similarly, if the destination of a stop on the return trip (Type 6 record) could not be identified by the reviewer, all Type 6 records would be marked with an "X" in the DLTFLAG field. Distance would be computed on the return trip for only the destination to origin records. This reduced the need for re-sequencing segment records associated with the trip.

(3) <u>If a foreign country destination was not reported or could not be identified, the following procedures were used.</u> The unknown destination was assigned to a major city in that country or to a city in the center of the country. If neither the city or the country could be identified, the corresponding trip records would be marked with an "X" in the DLTFLAG field.

• Cases with incorrect mode

If the mode for a trip or a segment of the trip was impossible (e.g., highway from United States to Europe), the record was edited to record the correct mode.

B. Non-highway Processing

Trip records sent to the computer-aided air/train processing system were reviewed as they were processed. The system was built with the "human-in-the-loop" capability which paused for operator intervention as unexpected situations were encountered. More detailed description on this system will be given later in Section 5.3. Most of the missing airports or train stations in segment records were identified with this computer-aided system. Editing on these airports or train stations was automatically made during this stage, when possible. Operators who performed this task were instructed to adopt the new speedy process rules when they encountered unknown

or missing values. All unknown locations that could not be easily identified during the "human-in-the-loop" process were marked with "X" in the DLTFLAG field.

4.3 FOLLOW-UP EVALUATIONS

Results from the speedy process were reviewed after a few Wave-Cycle data sets were completely processed. It was determined that certain modifications would be necessary in order to reduce the number of records marked with "X".   Trips that became out-of-scope (i.e., less than 100 miles each way) due to the X-marked segments, in particular, were reviewed further. Trip segment records were edited when possible and reprocessed to obtain valid segment distances. Total distance for each of these trips was recalculated, and the out-of-scope flags were then readjusted. This revised speedy processing procedure was employed to process all the remaining Wave-Cycle data files.

A brief study of the potential impact on the use of different levels of editing was conducted after operations on all ATS Wave-Cycle data files were completed. This study was based on two sets of Wave-Cycle ATS data. Results from this exercise are presented in Section 7 of this document.

5. ESTIMATION OF DISTANCES FOR TRIP SEGMENTS

5.1 INTRODUCTION

This section describes the methodology used to estimate segment distances for reported trips. As defined in the ATS, each trip could have several segments associated with it. For instance, a reported air trip from New York to Los Angeles would have a segment to the airport (access link or Type 2 record), the segment between the airport in New York and the appropriate destination airport in Los Angeles (if no other stops were reported), as well as the return segment from Los Angeles to New York. Each trip segment was treated as a separate O-D pair in distance calculation. The access segment was usually treated as highway, unless reported otherwise. These Type 2 records were processed using the highway distance estimation algorithm. The distance for the segments between airports was estimated using the air procedure. As another example, an auto trip from Washington, DC, to Knoxville, TN, might have reported stops in Roanoke, VA, and Kingsport, TN. Distance would be computed for the Washington, DC, to Roanoke, VA segment, the Roanoke to Kingsport segment, and the segment from Kingsport to Knoxville. The return portion of this trip would be computed in a similar fashion.

Approximately 70 percent of the total trip segments in the ATS required highway distance calculation. Of the remaining trip segments, the majority were by air. Trip segments by train made up less than two percent of total segment records, and water trips were less than one percent.

5.2 HIGHWAY DISTANCE

5.2.1 Highway Network Access

The national highway network is modeled as a collection of connected *links* and *nodes*. Links represent continuous portions of sections of highway. They are bounded by nodes which represent locations where some change in the movement of passenger traffic can occur, such as changes in speed or direction. Examples of nodes include highway intersections and interchanges, water ports, and airports. An important attribute of nodes is their geographic location, which is given by their latitude and longitude. Thus, nodes serve to define the shape of the highway network and play an important role in determining whether or not a highway link (i.e., roadway) is directly available to a well-defined geographic zone such as a zip code area.

The first step in finding a route for an auto trip segment involves determining whether the highway network is accessible at either the trip origin or destination, or both. If the network is accessible, the next step is to determine at which point the trip segment enters the highway network and at which point it exits. Because one may not live near a node on the network, some distance may be involved in

accessing the highway network at either or both ends of the trip segment. This distance is known as the *local access distance* and is represented in the network model as a *local access link*. Local access links model highway facilities not included in the national networks, such as local streets and roads. In graphical representations of a network, they are usually displayed as a line with one end attached to the node closest to where the trip segment enters or exits the network and the other end at a point representing the population centroid of the origin or destination zip code.

The national highway network represents approximately 400,000 miles of major roadways (see Figure 3). Not included in the network are an additional 3.5 million miles of urban minor arterials, major and minor collector roads, and paved and unpaved local streets, roads, and trails. Because of the existence of these minor arterials, collector roads, and local streets, the national highway network may be accessible to a zip code even if it does not physically pass through a zip code area. There are, however, a few zip codes which do not have access to the national highway network because the network does not include all local roads. Examples include a number of zip codes in Alaska.

For ATS, trip segments that begin by highway are assumed to start at the origin zip code population centroid and are assumed to terminate at the destination zip code population centroid. Highway access links, therefore, simulate the movement of vehicles over local streets and roads between the geographic center of the zip code area and nearby nodes on the national highway network.

The procedure used to generate highway access links to and from each zip code centroid is illustrated in Figure 4. The area around the centroid is divided into four quadrants. In each quadrant the national highway network node closest to the centroid within a 200-mile range is located. The straight-line distance d_i from the centroid to the closest node in quadrant i is calculated from the latitudinal and longitudinal coordinates of the two points. The straight-line distance from the centroid to the closest national highway network node in any direction is defined as d_{min}. Thus, each zip code centroid can have between one and four highway access links unless the closest highway node is over 200 miles away. In that case, the zip code area does not have access to the national highway network. The length of the access link in quadrant i is the straight-line distance d_i multiplied by a circuity factor of 1.2.

In Figure 4, the highway nodes closest to the centroid in each quadrant are labeled 1, 2, 3, and 4. Access links are created between the centroid and nodes 1, 2, and 4. No access link is created between the centroid and node 3 because the latter's straight-line distance from the centroid is more than three times the straight-line distance from the centroid to node 2, the closest of all highway nodes to the centroid.[5]

---

[5]Bronzini, M.S., Chin, S-M, Liu, C., Middendorf, D.P., and Peterson, B.E. *Methodology for Estimating Freight Shipment Distances for the 1993 Commodity Flow Survey*. Report prepared for the Bureau of Transportation Statistics and Bureau of Census, by Oak Ridge National Laboratory, Oak Ridge TN 37831, April, 1996.
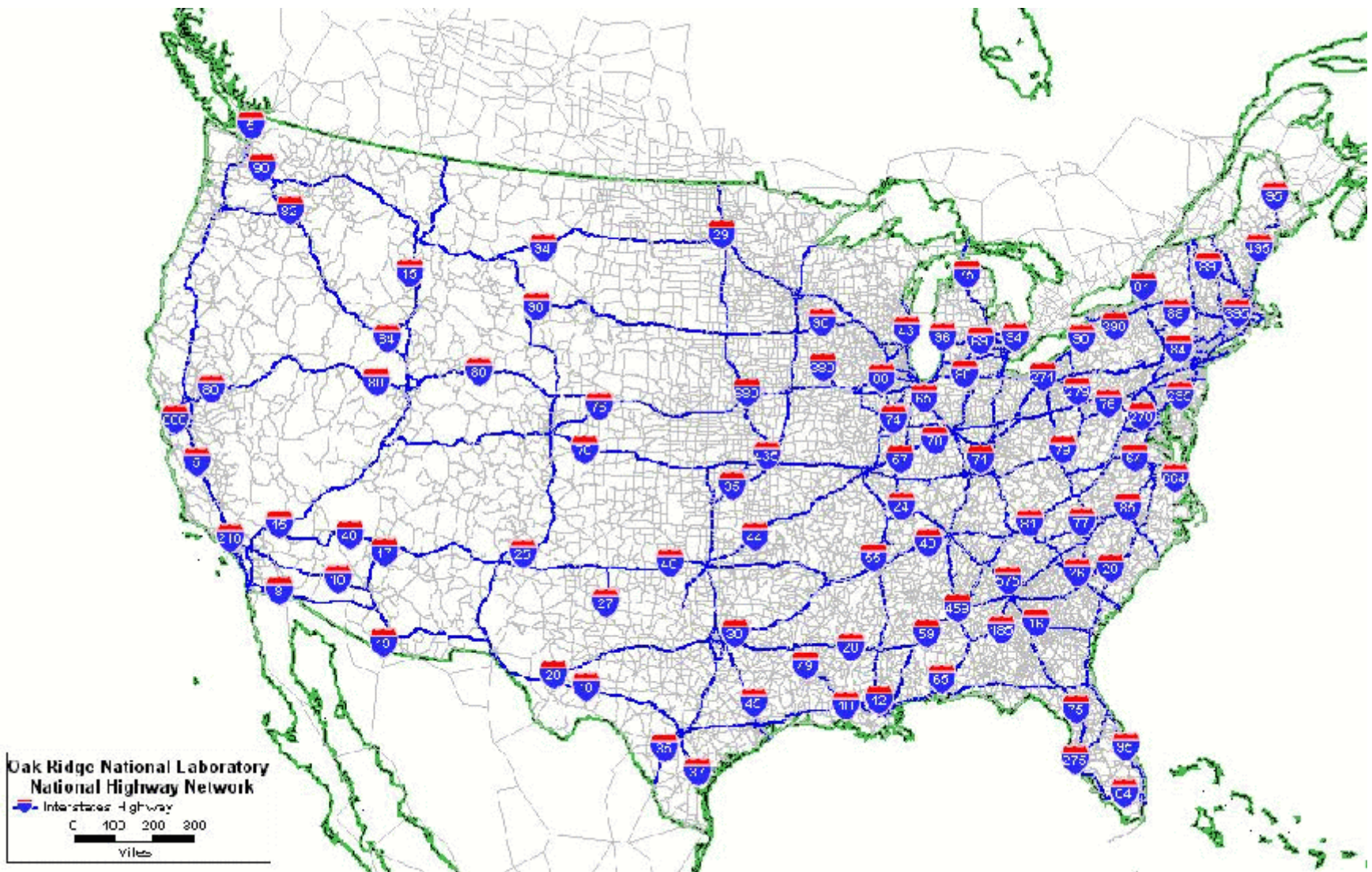
Figure 3. National Highway Network

49
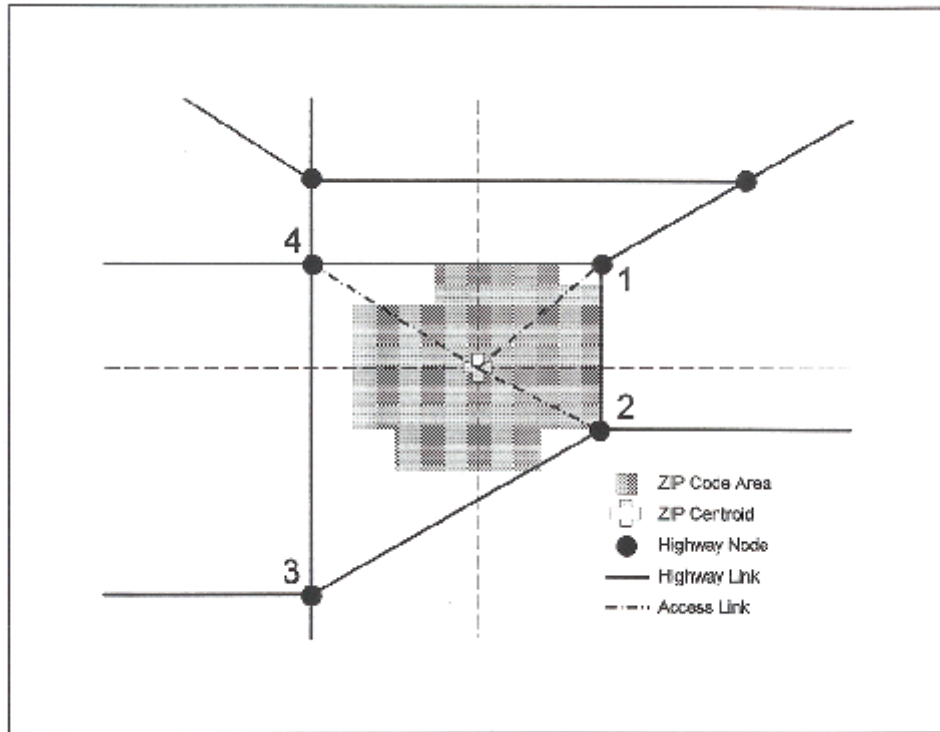
Figure 4. Determining Highway Access Links from a Zip Code Centroid

5.2.2 Highway Link Impedance

Simulating a route between any two points on the highway network is equivalent to a problem of finding the path with the least total impedance. Impedance, the generalized "cost" or penalty for using any link in a path, is measured in units that are close to "minutes" for clarity of presentation.

However, even though dominated by time, impedance also includes several other considerations and adjustments. For instance, it is commonly believed that travelers will tend to use freeways and interstates even when doing so will result in a route that is slightly longer in time than an alternative surface route. This may be due to considerations of safety, ease of travel, or lack of familiarity with local roads. To capture this effect, impedances for freeways were chosen to be slightly lower in the "minutes" units than actual travel times relative to other roads. In all cases, the point of an impedance function is to reproduce as closely as possible the paths taken by typical travelers, and it is not to estimate real speeds, times, or costs, however defined.

Impedance is of necessity a function of the known characteristics of the road. An effective speed is first determined for the road according to Table 8.[6] Suburban and urban ratings of the degree of urban congestion are made subjectively during network editing. Roads which are rated "rural", but are nevertheless within an urban area boundary, receive speeds from the "Fringe" column.

Table 8. Default Highway Speeds (mph)

Median	Access Control	Rural	Fringe	Suburban	Urban
Divided	Full	67	66	63	48
Divided	Partial	61	60	57	34
Divided	None	56	55	52	36
Undivided	Full	58	56	52	24
Undivided	Partial	50	48	44	27
Undivided	None/Multi-Lane	48	46	40	17
Undivided	None/2-Lane	45	43	37	13
Undivided	Unpaved	12	n.a.	n.a.	n.a.
	Ferry	12	n.a.	n.a.	n.a.

n.a. - not applicable

Table 9 shows the adjustments or penalties (determined by expert judgment) made to the default speeds in Table 8 based on roadway characteristics.

Table 9. Adjustments or Penalties to the Default Speeds

Roadway Characteristics	Adjustment or Penalty
Interstate	+8%
Principal Arterial	+1%
Collector (or below)	-1%
Toll Road	-4%
Winter Closure	-15%
Sub-Normal Pavement	-25%

[6]Southworth, F., Peterson, B.E., and Chin, S-M. *Methodology for Estimating Freight Shipment Distances for the 1997 Commodity Flow Survey*. Report prepared for the Bureau of Transportation Statistics and Bureau of Census, by Oak Ridge National Laboratory, Oak Ridge TN 37831, October 1998.

(The "sub-normal" rating is earned by narrow lanes, no shoulders, or unusually heavy grades or curves, to the extent these can be estimated from maps.) The resulting impedance for a link is its length divided by effective speed. (See Bronzini et al., 1996, for further details on the impedance calculations.)

5.2.3 Highway Routing Model

The following steps were used to determine the distance between two zip code population centroids (O-D pairs for each segment) on the U.S. National Highway Network.

Step 1.   If the national highway network is not accessible at either the origin or the destination zip code, report a distance of -1 to indicate that transportation by vehicle was not possible between the two zip codes. ORNL used several flags to indicate the type of problems encountered during the highway distance calculations. These flags are listed in Table 10.

Table 10. Flags Used to Indicate Problems Calculating Distances

Flag Codes	Descriptions
-1	Invalid Origin Zip Code
-2	Invalid Destination
-3	No Access at Origin
-4	No Access at Destination
-5	Unreasonable Mode
-6	No Access/Unreasonable Mode

All trip segment records with flags coded (i.e., other than blank) were sent back (or extracted) for review and proper editing. Those without problems from Step 1 proceeded with the routing process.

Step 2.   The highway distance calculation algorithm finds the minimum impedance path between the two centroids on the national highway network, including the highway access links at the origin and destination.

Step 3.   The highway distance for the given trip segment is calculated as the sum of the link lengths on the selected path.

Step 4.   For access segments (i.e., Type 2), if the GCD is 10 miles or more, then the procedure described in Section 5.2.1 is used. However, if the GCD for this segment is less than 10 miles then:

(a)   Compute a non-network-based highway distance equal to 1.2 x GCD. This represents the distance along a path that uses only local streets and roads not included in the national highway network.

(b)   Find the minimum impedance path between the two centroids on the national highway network, including the highway access links at the origin and destination. Compute the network-based highway distance as the sum of the lengths on this path.

(c)   If the network-based distance is more than 1.5 times longer than the non-network-based distance, use the latter as the highway distance for the zip code pair. Otherwise, use the network-based distance.

(d)   In the rare event that the GCD is 0, the origin or destination zip codes are the same or one of the zip codes is a point location embedded within the other zip code. In this case, because average distance between any two points of a circle is 0.91 times the radius of a circle,[7] the distance of this segment is, therefore, estimated as 0.91 x R_e x 1.2 or 1.1 x R_e. In this calculation R_e is defined as the radius in miles of a circle whose area equals that of the enclosing zip code, and 1.2 is the assumed circuity factor.

5.3 DISTANCE ESTIMATION FOR AIR TRIP SEGMENTS

The length of a given trip segment when traveled by air was measured by the GCD between the two airports identified for each end of this segment. Information on the airport which the respondent used to begin his/her trip was collected in the ATS. This information was stored with the Type 2 records in ATS data files. ATS also asked the survey respondents to report locations (city and state) where a change of transportation modes or a transfer between airplanes had occurred. In most cases, the information provided could be used directly to locate the airports, and the GCD for each of these trip segments could then be computed based on the latitudes and longitudes of these airports. If no airport information was given for the return trip, and if air was specified as the mode of transportation for the return portion of the trip, ORNL assumed the same airports as those used for the departure and calculated the distance accordingly.

---

[7]Kendall, M. G., and Moran, P. A. P., *Geometric Probability*, Griffins Statistical Monographs and Course No. 10, ed. Kendall, London, 1963.

When the actual location of an airport utilized by the traveler was not explicitly identified, a computer-aided searching routine developed by ORNL was used to identify the most likely airport location so that distance calculation could be performed. Furthermore, air trip segments with missing or misspelled names (city or airport), or with obviously misplaced geography (e.g., wrong state) were also treated by this computer-aided program to find the most reasonable airport location and to estimate distances for these air segments.

The air trip routing process located the most likely airports by hub size ranking (large, medium, small, and non-hub) and by their distances from the given city (i.e., the end point of an air segment). Larger airports were put at the top of the list with proximity to the population centroid of the given city as the secondary ranking criteria. In the cases of foreign trip segments, no attempts were made to locate airports in the foreign country. Geographic location (i.e., latitude and longitude) of the foreign city was used in calculating GCD for these air segments.
Alaska airports proved especially challenging and often required manual intervention to select applicable routes.

5.4 TRAIN ROUTING

A small percentage of the total number of ATS trips segments were by train. The passenger rail network used for domestic train trips in the ATS was an Amtrak passenger rail network and its connecting bus service. It was based on two network databases developed and maintained by the Federal Railroad Administration (FRA): the Railroad Network Database and the Amtrak Stations Database.

The FRA Railroad Network Database was digitized and organized according to each individual railroad company. Since trackage rights for Amtrak are flagged in the link data layer, the Amtrak network was created by extracting all links that were flagged as being used by Amtrak. ORNL made modifications to this network so that routing could be performed.

The Amtrak Stations Database is a point-based facility databases digitized independently of and with no logical connection to the Railroad Network Database. Thus, ORNL transcribed Amtrak Stations Database information to the Railroad Network Database. Editing was also done to transcribe the FRA Amtrak Station name to the ATS Amtrak Railroad Network Database.

The most up-to-date passenger service schedule was obtained from Amtrak, and editing was done by ORNL on the ATS Railroad Passenger Network to reflect the most up-to-date national railroad passenger service. The origin and destination of each rail-related trip segment were identified by the Amtrak station names. Station names were used to identify the origin and destination nodes (latitude and longitude) within the ATS Passenger Railroad Network Database. A path based on the Amtrak published schedule was found, and distance was then computed for the path.

A few of the trip segments were foreign segments using train as their main mode of transportation. Distances for these trip segments were calculated based on GCD between the segment origin and destination.

5.5 WATER DISTANCE

Only a very small percentage of trip segments in the ATS were reported using ships or boats as the mode of transportation. Most of these involved cruise ships or travel within foreign countries. In the case of cruises, it was almost impossible to determine the distance traveled by water without spending much effort for little gain. For all water-related segments, therefore, estimated distances were generated using GCD x 1.2.

6. QUALITY ASSURANCE (QA) /QUALITY CONTROL (QC) PROCESS

The computer programs that ORNL developed to compute trip segment distance also performed QA and QC checks on the data. These checks were performed in addition to the pre- and post-processing QA/QC activities. All QA/QC processes were conducted in order to ensure the accuracy of the data as well as the results. The Pre-processing QA/QC procedures ensured that data errors such as missing or miscoded locations, invalid modes of transportation, or out-of-sequence records were corrected, if possible, before the distance calculation routine began. The post-processing QA/QC processes further checked trip records to validate results. More detailed discussions of these activities are included in Section 3. Certain data errors were difficult to detect during the pre-processing QA/QC process. The miscoding of zip codes when first encountered (i.e., prior to inclusion in the ZIPFIXED database) was one example. Unreasonable use of a given type of transportation which was unaccessible at the location was another example. Consequently, data checking was also important during the distance estimation stage.

The checking process within the distance estimation phase was an iterative operation. If the distance calculation programs detected a problem in a trip segment, this record was returned with a specific negative value in the flag field (see Section 5.2.3). Whenever possible, records with negative flag values were checked, edited (if necessary), and re-processed through the distance programs. Additional accuracy checks were performed on the distance results. Trip segments identified as having problems were either edited and re-run through the distance calculation programs or marked for deletion if corrections could not be made with limited efforts (e.g., more than 10 minutes).

The following sections describe these QA/QC checks performed during the distance estimation operation. Brief discussions on how problems were handled under the original and the speedy processing procedures are also presented.

6.1 QA/QC DURING THE HIGHWAY DISTANCE CALCULATION PROCESS

A sequence of four programs was used to process ATS records where the trips or trip segments utilized highway mode. ATS highway mode included use of private automobile (car or truck), rental vehicle, taxi, motorcycle, recreational vehicle, and bus. A separate program was also developed to check access links for trips where the main mode of transportation was non-highway (or non-auto). These five programs are briefly described below. The first two programs were run as part of the pre-distance calculation effort. The third program was run during distance estimation, and the remaining two, as part of the post-distance estimation validation.

(1)   Foreign Country Destination Check

Because of an error detected early on in the access link procedure, an additional program was written to ensure that no records with a highway mode had a foreign country destination other than Canada or Mexico. This program was run prior to the distance calculation programs. Trip records with errors were returned to the editing stage for correction. Air was assumed as the correct mode for these trip segments unless otherwise indicated by other variables in the data file.

(2)   Consistency Check

The consistency check program compared the origin and destination zip codes in Type 1 (i.e., trip segment from home to final destination) and Type 7 (i.e., trip segments from final destination to home) records to ensure those zip codes matched. That is, it checked to see whether the origin zip code in Type 1 matched the destination zip code in Type 7 and whether the destination zip code in Type 1 matched the origin zip code in Type 7. The program also checked any record segments on the going and/or the return trip segment to ensure that the destination zip code of one segment matched the origin zip code of the subsequent segment. Any errors found were flagged, and records with these error flags were returned for editing. This program was run before the distance calculation programs.

The review actions under this process were conducted slightly differently for the original and the speedy processing procedures. Under the original processing procedures, zip codes where the origin and destination pairs did not match were corrected. Unreported stops or destinations were identified using the procedures described in Section 3. Under the speedy processing procedures, however, records found with errors/problems were not always corrected. Those that could easily be corrected (e.g., typos in zip codes, obvious misspellings of cities) were edited to the known zip codes. When zip codes could not easily be identified, these records were marked with "X" in the DLTFLAG field.

(3)   Connectivity Check

The connectivity checking program was run as part of the distance calculation program. It checked all zip codes to be sure that they matched existing zip codes in the zip code database. It also checked to ensure that there was a highway connection between each origin and destination pair. Any records that did not match existing zip codes were marked with a "-1" (no origin zip code found) or a "-2" (no destination zip code found) value in the flag field. Any records that had no highway connection were marked with a "-3" (no access at origin) or a "-4" (no access at destination) value in the flag field.

The review actions under this process were conducted the same way for the original and the speedy processing procedures. Typos in zip codes were corrected. The mode was edited for

those trip segments where there was no highway connection. Typically, for these cases, the mode was assumed to be air, unless the trip information indicated otherwise

(4)   Circuity Check

The circuity check program compared the highway algorithm distance result from origin to destination against the GCD from origin to destination. As a rule of thumb based on transportation specialists' expert advice, if the distance result was three times greater than GCD (GCD x 3), then the trip was flagged for further verification and checking.

The review actions under this process were conducted slightly differently for the original and the speedy processing procedures. Under the original processing procedures, each trip with a distance result greater than GCD x 3 was reviewed using the commercially available software *Map & Go* to route trips for distance verification. In most cases the distance estimates of the trips were consistent. The stops along the way generally accounted for the distance discrepancy. In a few cases, the stops were out of sequence. These trips were reordered and edited with new stop sequence numbers and re-submitted for distance calculation. Under the speedy processing procedures, however, trips with a distance result greater than GCD x 3 were, in general, not individually reviewed because of time factors. Only those trips with unusually large distance discrepancies (where computed distance was greater than GCD x 10) were reviewed, edited, and re-submitted for distance calculation, when necessary.

(5)   Distance Check

The distance checking program checked the distance result added from each segment on the going trip against the direct distance from origin to destination summed by the highway algorithm. It also performed similar checking for the return portion of the trip. If the distance result from summing the segments of the added stops on the going or the return trip was three times greater than the GCD distance measured between origin and destination, then the trip was flagged for further checking.

The review actions under this process were conducted slightly differently for the original and the speedy processing procedures. Under the original processing procedures, each trip where the distance sum of all stop segments on the going or the return direction was 3 times greater than the direct distance was verified using *Map & Go* software. In most cases the trip distances were consistent. The stops made along the travel path accounted for the distance discrepancy. In a few cases, the stops were out of sequence. These trips were reordered and edited with new stop sequence numbers, and the distance was re-calculated. Under the speedy processing procedures, however, these cases were, in general, not reviewed because of time factors. Only those trips with unusually large distance discrepancies (where computed distance was greater than GCD x 10) were reviewed.

A final, manual check was made of the file that contained distance to ensure that the records, etc., in this file matched those in the original file without the distance result. The purpose of this check was to ensure that distance results were merged with the correct records.

6.2 QA/QC DURING THE AIR AND TRAIN DISTANCE ESTIMATION PROCESS

A series of computer programs was developed to process trip records that included air or train as the mode of transportation for at least one of its segments. These trip segment records were processed slightly differently depending on their types. Consistency checking and distance calculations were performed on all trip types. Trips that included stops or side trips required further consistency checking and distance calculations on those additional segment records, as well as needed special handling of possible mode changes between stops or side trip segments. As defined in the ATS, commuter trips and mover trips are one-way (see Section 3.3). Therefore, they were treated separately, and fewer programs were called to process these types of trips.

An important part of the air and train data processing was performing two activities: (1) the identification of airports and train stations and (2) the geocoding of these airports and stations. Furthermore, highway segments which connected airports or train stations to zip code centroids of given place names were added to the trip records when appropriate. Since zip code centroids were not likely to be located at the airport or train station, such segments were needed to represent access links to or from the airports or train stations. Note that transfer between two flights or trains did not require an access link because distances for air and train segments were calculated based on the geographic locations of the airport or train station (i.e., latitudes and longitudes). For instance, suppose a trip began from one's residence in Boston, included a flight out from Logan Airport, a transfer in Atlanta, and continued to Orlando on another flight. The trip segment between home and the Logan Airport is the "access link" connecting the zip code centroid to the airport. If the trip records were reported and coded correctly, this segment should be reflected in a Type 2 record. The distance between Logan and Atlanta would be calculated based on the two pairs of latitude-longitude of the two corresponding airports. An access link would be added to connect Orlando Airport to the zip code centroid of the city of Orlando. Access links were generally assumed to be highway segments.

The distance estimation process for air or train segments utilized a unique computer-aided, with human-in-the-loop, system design. This design was intended to capture data errors and correct them, if possible, as soon as they were identified. If an operator, the human-in-the-loop, could not correct a specific data problem as it was identified, the option to flag and skip that record for later checking and editing was also provided by this computer system. The rationale for this design was that by allowing human intervention when errors were encountered during the automated processing, the efficiency of the system and the data quality would be improved.

Under this system design, trip records were processed by a modular approach. These modules are briefly described in the following sections.

6.2.1 Consistency Checking for Zip Codes

A.   Type 1 vs Type 7 Records

As described in Section 2.1.2, a Type 1 record is the main trip record that holds information about the out-bound trip away from home to the final destination. A Type 7, on the other hand, holds trip characteristics associated with the return direction of that trip. By ATS's trip definition, the origin zip code in a Type 1 record should be the same as the destination zip code in a Type 7, and vice versa. This consistency checking was performed to ensure no coding errors or missing data were contained in these two record types. The mode of transportation contained in these two records did not necessarily have to be the same, however. Although in most cases the main mode used for both travel directions in a trip were the same, many trips did utilize different transportation systems to complete their travels. When the mode of transportation was unreported (i.e., missing) in a Type 7 record, the system assumed that same mode as indicated in the associated Type 1 record was used.

B.   Type 2 Record

Consistency checking for the Type 2 record was conducted to ensure origin place name and zip code specified in this record were the same as those specified in its Type 1 record (in most cases, this was the location of residence for the household). If a Type 2 record was not presented, this procedure also assured that no Type 3 record was included as a segment in the given trip. Missing airports or train stations from the Type 2 records were handled by a module described under 6.2.2 of this Section.

C.   Other Record Types

All other record types, with the exception of commuter trips, which contain only one record per trip, were checked to ensure each origin and destination place names and zip codes were in logical order. That is, the origin from one segment record should be the same as the destination specified in the segment record before it; and the destination from this segment record should be the same as the origin of the following segment. Editing procedure for errors found from this and other checking processes followed the same rules and steps as described in Sections 2 and 3.

6.2.2 Airport and Train Station Identification and Geocoding Module

As mentioned earlier, great effort was placed on identifying airports/train stations and geocoding of airports/train stations during the processing of ATS air and train trips. Two airport databases as well as

an Amtrak station database were used by this module. The airport databases included one for about 400 major commercial airports and another for all other airports and airfields. These databases contain facility names, cities and states where they are located, and geographic coordinates (i.e., latitude and longitude) of these transportation facilities. For the airport databases, hub size is also included. Similar to other databases used in the ATS data processing, various spellings and new items found that could be reused for ATS were added to the associated databases so that later occurrence of the same types of errors would be automatically corrected by the computer system.

Again, since zip code centroids for most of the U.S. cities are not located exactly at airports or train stations, it was necessary to use access links to simulate the travel segments between residences (or other places) and the transportation facilities such as airports and train stations. If a Type 2 record was not included as the access segment of the out-bound trip, one was created to connect from the zip code centroid of the place where the air or train trip started (i.e., where the respondent transferred from auto to air or train travel) to the nearest major commercial airport or Amtrak station identified by the computer program. Transfer between two flights or trains does not require an access link. Travel distances for these segments were calculated from latitudes and longitudes of the two airports, or Amtrak stations, at the two ends of these trip segments.

The human in-the-loop design was the key component in this module. This unique design was especially useful when a variation of airport names was encountered during the processing of air segment records. Operators of the system could select from a separate display window of airports to identify the correct spelling of airport names. They had the option to save this information to the airport databases for future use. Moreover, information displayed in this window could appear in different order. If the zip code of an airport or place name where this airport is located was provided, airports would be displayed in the order of hub sizes and their distances from the given zip code or place. Otherwise, airports would be sorted by hub sizes and state names. Generally, this search started with the major commercial airport database. The system operators could specify the use of the auxiliary airport database, especially when handling cases utilizing personal aircraft, military airbases, or airports located in the state of Alaska. In addition, if on-the-spot corrective actions for problems found during the computer-aided data processing operation were difficult to determine, the operators had options to instruct the computer system to skip records for follow-up review or editing at a later time.

The system also allowed operations to be interrupted in mid-stream of the data processing. When an operation was resumed later, the system automatically began its process from the first record it encountered which was not already geocoded. It was also possible to specify a record number, if preferred, for the system to begin its execution for airport identification and geocoding.

A similar procedure was used to process trip segments utilizing Amtrak. The Amtrak station database is much smaller in size due to the limited size of the Amtrak network. Bus connection provided by Amtrak as part of their service was also included as part of the Amtrak network. Those stations, therefore, were considered as part of the Amtrak stations.

Foreign airports and train stations were geocoded at the place name level. No attempt was made to identify geographic coordinates for airports or train stations located in foreign countries. In a few rare cases, survey respondents reported using Canadian airports near the U.S. border to begin their travels. Geographic coordinates of these Canadian airports were later added to the airport databases.

6.2.3 Data Handling and Distance Calculation for Air or Train Trip Segments

After the geocoding of airports and train stations was completed for all record types, distances for all air and train segments were mathematically calculated as described in Section 5. For segments that began with an airport and ended at an airport, air distance was estimated as the GCD between these two airports. Both network-based distance and GCD were estimated using the Amtrak network database for the train segments. If a Type 2 record was not included for a given air/train trip, a highway access record connecting the zip code centroid of the departing city to its airport (or Amtrak station) was added prior to the first air/train segment. Similarly, a highway egress record connecting the arrival airport/train-station to the given city was also added. These access or egress links were assumed to be highway links. These link records were needed so that mode-specific distances could be correctly assigned for each ATS trip. Along with highway trip segments, these link records were transmitted to the highway distance calculation process, and highway distances were estimated.

6.2.4 Trip Segment Linkage and Final Trip Distance

When transmitting trip segments between various processes, all record identification information was carried with a segment so that individual trips could be reconstructed at the end of the ATS processing. All highway distances calculated for a trip were merged back with their associated air or train segments to obtain the total trip distance. Round-trip distances, both GCD and network-based distance estimates, were added to the Type 1 record for each ATS trip. To determine whether a reported trip was in-scope for the ATS purpose, network-based one-way distances for both out-bound (i.e., from home to final trip destination) travel and in-bound (i.e., from final destination back to home) travel were also calculated. These "directional" total distances were added to record Types 1 and 7, respectively. If both one-way trip-lengths for a given trip were found to be less than 100 miles, a '9' would be coded for the DLTFLAG field in the Type 1 record. Although not explored under this project, it was anticipated that these directional distances could be useful in studying passenger travel behavior or mode choices.

In addition to the above-mentioned distances, total international and total domestic travel distances were also summarized separately and reported by ORNL in two new variables added to the Type 1 records. Travel outside the 50 U.S. states and DC, including U.S. territories and Puerto Rico, was considered to be international under ATS. Highway and train travel were divided at the U.S.-Canada and U.S.-Mexico borders. Distances traveled on highways within the U.S. borders, within Hawaii, and within Alaska were all considered domestic. Air trip segments that fell entirely in a foreign country,

connected the last U.S. exit airport to a foreign location, or linked from a foreign location to the first U.S. entry airport, were all regarded as international segments.

6.3 FINAL QA/QC PROCESS AND RESULT TRANSMITTAL TO CENSUS

After all ATS trips were processed, edited, and mode-specific distances were added to trip records, a temporary variable representing a ratio between the total round-trip network-based distance and the total round-trip GCD was created. This ratio is also known as the circuity factor for a given trip. Distribution of circuity factors was then generated and extreme values were then reviewed. The final QA/QC process also included a quick check of the DLTFLAG fields to ensure flags were properly identified and those marked as out-of-scope trips (i.e., less than 100 miles one-way) were correctly classified. Any trip records that required re-editing and re-processing were sent back to the appropriate processing step, and all procedures as described in this report were repeated.

Recall that ATS data was processed as Waves and Cycles. At any given time, different Wave-Cycle data sets could be undergoing different data processing stages. All trip records under the same Wave-Cycle were placed back into a single SAS data file after all processes were completed. This data set was then converted to the SAS transportable file format before copied to a computer tape. As described in Section 1, all computer tapes were hand-carried by an authorized ATS SSE to a local Federal Express office and delivered to the Census.

7. IMPLICATIONS OF THE QA/QC PROCESSING METHOD

7.1 INTRODUCTION

Initially, it was both BTS's and ORNL's intention to save as many ATS trip records as possible from being eliminated because of problems in the data. Thus, during the processing of W1C1 data, ORNL utilized many tools (e.g., hardcopy atlases, phone directory CD-ROMs, geographic software products, Internet searches, and other reference materials) to locate misspelled or uncoded place names and zip codes as well as to resolve problems with particular cases. A draft detailed data processing procedures document was assembled over the first few months of the W1C1 data processing period. This served as a living document to guide the data processing activities in subsequent Waves and Cycles. It also specified the database structures and record layouts for files being transmitted between ORNL and the Census. The data processing procedures document was updated frequently as new problems and their solutions were identified.

During the W1C1 data processing period, extensive manual processing was routinely conducted. This included activities to locate correct zip codes, to obtain latitude-longitude for foreign place names, as well as to examine reasonableness of the mode and sequencing of the stops as reported by the respondents. Three databases were compiled based on results gathered from these activities: one for latitude-longitude of foreign place names (ATS_FC); another for domestic place names that were miscoded as foreign country locations (REALYUSA); and a third one for domestic zip code fixes (ZIPFIXED). These databases were updated regularly throughout the life-cycle of the ATS data processing as new problems were encountered. At the conclusion of the ATS data processing, the ATS_FC database file contained a total of approximately 12,000 records. The REALYUSA data set had 33 place names and the ZIPFIXED file included over 3,100 entries.

With knowledge learned during the processing of W1C1 data, a computer program called AUTOFIX was developed to utilize these three databases so that problems found in all subsequent Waves and Cycles could be systematically fixed. AUTOFIX also edited the mode of transportation for trips in which respondents reported an impossible mode, e.g., highway travel between the Hawaii islands or between a domestic location and a foreign country other than Canada or Mexico. These trips were generally changed from highway to air unless another mode of transportation was indicated.

Not all of the problems encountered by ORNL could be systematically repaired by the AUTOFIX program. Problems found in later Waves and Cycles might not have been observed in the earlier Cycles. Also, many problems were trip specific and required a closer review of records before corrections could be made. For example, when respondents provided rather descriptive information (e.g., "10 miles from city X" and no information about the state, or "stopped along Rt. 123") a closer examination of the trip records would be required to assign the zip code needed to calculate the trip distance. Furthermore, resequencing of the reported stops (e.g., when they are clearly out of sequence

based on the other information provided in the trip record) could not be done by a totally automated procedure. ORNL would review the detailed information included with this trip record in order to resolve these types of problems.

Obviously, extensive manual data checking and editing is a time-consuming process. During the latter part of Cycle 1 data processing, BTS, ORNL, and Census met to discuss ways to accelerate the data checking and editing procedure. It was determined that such a detailed editing process might not be cost effective, thus it was decided to eliminate all manual editing and use the automated editing procedure alone, i.e., the AUTOFIX program. Deletion flags were used to mark those cases in which distance calculations could not be completed so that Census could eliminate those records from the final ATS file. Unfortunately, use of the automatic process alone led to the elimination of more records than was deemed acceptable by the ATS team, thus ORNL adopted a modified approach (see Section 4) in which data were initially processed using the AUTOFIX program followed by a method referred to here as the "manual quick-fix," a quick manual screening of the data to identify and correct records that could be saved with minimal manual editing. This approach was then applied to the processing of all data files after Cycle 1.

7.2 DESCRIPTION OF THE DATA SETS

The original trip files transmitted by Census to ORNL for distance calculations contained records for all trips collected during the ATS interviews. Trips reported as identical, however, were only included once in these files. After ORNL processed the trip files and returned them to Census, Census used the calculated distances to create replicate distance estimates for the reported number of identical trips. All records that were flagged as short trips (less than 100 miles one-way) or marked for elimination due to insufficient information were excluded from the final data set. Adjustments for each record were estimated by the Census to generate expansion factors for both household trips and person trips. The resulting final trip files generated by the Census, therefore, do not contain the same number of trip records as those in the original trip files.

The study reported in this section was conducted after all data processing activities at ORNL were concluded. It evaluates the impact of ORNL's QA/QC processes on the entire ATS data set as well as on two subsets, W1C1 and W1C3. The original ATS trip file received from Census and processed by ORNL contained records for approximately 294 thousand trips, which included over 33 thousand trips from W1C1 and about 26 thousand from W1C3. The total number of round trips in Census's ATS final household trip file is approximately 338 thousand, which expands to about 685 million household trips. The final household trip file consists of approximately 43 thousand W1C1 round trips, which represent a total of over 69 million household trips. The final household trip file also includes records for approximately 26 thousand W1C3 trips, representing about 58 million household trips. That is, data for about 10 percent and 8 percent of total household trips were collected during W1C1 and W1C3 periods, respectively. The final person trip file consists of 556 thousand records

representing over one billion person trips. Approximately 68 thousand records representing 107 million person-trips were from W1C1; approximately 41 thousand records representing 83 million person-trips were from W1C3 (see Table 11).

Table 11. Number of Trips (in thousands)

	Number of trips (unweighted)[1]			Number of trips (weighted)		
	All	W1C1	W1C3	All	W1C1	W1C3
ORNL ATS trip file[2]	294	33	26	na	na	na
Final ATS household trip file	338	43	26	684,662	69,378	57,782
Final ATS person trip file	556	68	41	1,042,615	107,053	83,674

[1] Does not include records for identical trips.
[2] Each record in these files represented a complete round trip from origin to destination and back.

7.3 EVALUATION OF THE EDITING PROCESSES

The major objective of this evaluation study was to determine the impact of ORNL's QA/QC processes on the quality of the ATS trip data. This study focused on the following issues:

- determining the overall impact of ORNL's editing on the ATS data;
- comparing results generated by the original detailed processing method to those generated by the revised procedures implemented later during ATS data processing;
- assessing the effectiveness of the AUTOFIX program in systematically correcting data errors or imputing missing values; and
- determining the benefits of using the "manual quick-fix" procedure.

Some of the recommendations for the processing of ATS data in the future that are included in Section 8 of this report were based on the findings from this study.

7.3.1 Overall Impact

Approximately 20 percent of the person-trip records (108,761 records), or 24 percent of the household trip records (79,701), were modified by ORNL during the editing process. Each of these records could have one or more fields in it that were edited. The editing process actually affected slightly higher proportions of the weighted data; 21 percent of the person-trips or 25 percent of the household trips were affected by ORNL's editing procedures (Table 12). Changes most frequently involved identifying missing or incorrect location information (origin, destination, stop, or station zip codes), but errors in the reported mode of transportation were also corrected. The station variables were most commonly modified by ORNL because station geographic locations necessary for distance calculations were not provided by the Census. Editing was most common on longer trips as these trips involved more stops and more mode transfers. Consequently, trip records modified by ORNL during the editing process account for 54 percent of the total person-miles or 59 percent of the total household-miles (Table 13).

Table 12. Trips Affected by Editing (in thousands) - weighted

	# of household trips	% of household trips	# of person-trips	% of person-trips
Total (ATS)	684,662	100 %	1,042,615	100 %
Total edited	173,194	25 %	222,361	21 %
Total with edited location excluding station variable	61,009	9 %	82,907	8 %
Total with edited location including station variable	144,396	14 %	181,365	17 %

Table 13. Household Trip-Miles and Person-Miles (in thousands) - weighted

	# of household trip-miles	% of household trip-miles	# of person-miles	% of person-miles
Total ATS	730,803,347	100 %	1,043,180,926	100 %
Total affected by the editing process	428,221,034	59 %	563,384,798	54 %
Total with edited location excluding station variable	202,421,572	28 %	274,119,908	26 %
Total with edited location including station variable	395,635,034	54 %	517,525,436	50 %

7.3.2 Comparison of the Results

As mentioned earlier in this section, the objectives of this study were to (1) examine the effectiveness of the AUTOFIX procedure; (2) to determine the benefits resulting from the "manual quick-fix"; and (3) to compare results generated using the "detailed manual checking and editing" processing method vs. those generated using the "revised processing method" (i.e., AUTOFIX plus the "manual quick-fix"). The results of this study can provide useful information for the planning of ATS 2000.

7.3.3 Evaluation of the Benefits

Since the AUTOFIX program was built on experience gained through the detailed processing of W1C1, a study of its effectiveness based on the same data set will certainly produce biased conclusions. Furthermore, databases used by the AUTOFIX program were updated as new problems were identified over the duration of the ATS data processing. Thus, data sets from later Waves and Cycles of the ATS generally benefit more from and contributed less to the contents of the databases used in the AUTOFIX. Nonetheless, since the W1C1 data file was the only file processed entirely by the detailed method during the original ATS processing, this file was used to compare the results generated using the "detailed manual checking and editing" procedure to those generated by the "revised processing method".[8] The original W1C1 data file (as received from the Census) was re-processed using the revised procedures to generate a new result file. In other words, the W1C1 data set was processed with the systematic checking and editing program, AUTOFIX, and followed by the "manual quick-fix" (i.e., the quick manual screening and fix procedure). Result files generated from both the detailed and the revised procedures were then compared to evaluate whether the detailed checking and editing process results in substantial improvements in the data quality.

All W1C1 trip records edited by ORNL during the detailed processing procedure were extracted to form a subset. This subset was then merged with corresponding trip records from the W1C1 reprocessed file which used the revised procedure (AUTOFIX + "manual quick-fix"). The resulting data set, thus contained distance estimations calculated based on modifications done using both processing procedures.   This study data set was then merged with expansion factors obtained from Census's final household trip file to create a 12,445 modified trip-record file, NUW11CMP. This file includes roughly 29 percent of the W1C1 trip records. The study data set was also matched with expansion factors from the final person trip file to form another file, PRW11CMP. This edited person trip file contains 16,405 person trip records, accounting for roughly 24 percent of the W1C1 person trip file. Records included in both NUW11CMP and PRW11CMP files were either edited under both processes or by the detailed editing process only. It is possible the edits/changes made under the revised process might be different from those generated during the detailed editing process, because a

---

[8] It would be ideal if this comparison study could be conducted with a fresh set of trip files, but given the limitation on available resources, it was not possible to perform such extensive checking and editing procedures to other data sets for the purpose of this comparison study.

quick review of the data might result in a different "fix" than a more detailed examination. Without editing by ORNL, the trip records included in data sets NUW11CMP and PRW11CMP would either have been marked for deletion or estimates of trip distances would have been inaccurate.

In examining the NUW11CMP trip records, 624 of these 12,445 trip records did not have matching household trip records. Hence, no expansion factors were available for these 624 trip records.   A review of these 624 cases revealed that 434 records were for out-of-scope trips, which include 4 trips with foreign country origins and 430 short trips (i.e., less than 100 miles one-way). Another 8 of the 624 trips were found to have flags in the edited file indicating elimination due to lack of destination information. These 442 records were consequently excluded from Census's final household trip file. No reasons could be identified for the remaining 182 trip records, however, as to why they were excluded from the final household trip file. Over 75 percent of these 182 trip records were air trips and about 20 percent were auto trips.

After elimination of the 624 non-matching records, 11,821 trip records remained in the W1C1 study set. These trip records represent a total of 19 million household trips and roughly 25 million person trips (Tables 14 and 15). Thus, mode-distance measurements for about 27 percent of the total W1C1 household trips and 23 percent of total W1C1 person trips would either have been computed incorrectly or would not have been computed at all if this editing had not been done by ORNL.

Table 14. W1C1 Study Set

	# household trips	% of total W1C1 household trips	# person trips	% of total W1C1 person-trips
total in final W1C1	69,378	100 %	107,053	100 %
total in edited W1C1 (by the detailed method)	19,032	27.4 %	4,624	23.0 %
records either edited differently or not edited by the revised method	6,162	9 %	8,069	8 %
number deleted in the revised method but saved by the detailed method	475	0.7 %	678	0.6 %

Table 15. W1C1 Study Set - Trip-Miles and Person-Miles

	trip-miles	% of total W1C1 (trip-miles)	person-miles	% of total W1C1 (person-miles)
total in final W1C1	68,751,625	100 %	99,099,105	100 %
total in edited W1C1 by the detailed method	43,133,766	62.7 %	57,209,079	57.7 %
records either edited differently or not edited by the revised method	15,347,655	22.3 %	20,925,533	21.2 %
number deleted in the revised method but saved by the detailed method	572,534	0.8 %	810,254	0.8 %

In terms of data items that were modified, results from both the detailed and the revised processes generally agreed with one another. Of the 19 million household trips edited in the detailed processing method, 6.2 million trips would either have been edited differently or not edited in the revised method. Similarly, of the 24 million household trips edited in the detailed processing method, 8.1 million trips would either have been edited differently or not edited in the revised method. For these records, distances could have been calculated, but the distances would have been incorrect due to errors in mode or location coding. Furthermore, without detailed editing, roughly one percent of the total W1C1 records would have been marked for deletion in the revised processing method.

The mode of transportation was most frequently modified for air trips under both methods. About 99 percent of mode changes under the revised process were for edits and only 1 percent for imputation of modes. On the other hand, 52 percent of mode changes were for edits and the remaining 48 percent for imputation with the detailed processing method. This was because, in most cases, imputations occurred after additional reviews and investigations of the trip records were conducted. Generally during the systematic AUTOFIX procedure, (i.e., the main part of the revised process) and the minor manual editing that followed (i.e., quick screening + fix), missing zip codes or miscoded locations were identified and replaced with corrected information. Imputation of missing data items was rarely applied under the revised process.

7.3.4 Comparison of the AUTOFIX

The W1C3 data set was selected to determine the effectiveness of the AUTOFIX utility program as well as the added benefits from the "manual quick-fix". This set was selected because it was one of the middle sets and was also a smaller file. All trip records from the W1C3 result file that were modified

by AUTOFIX only as well as those with additional quick screening and fix were extracted. Expansion factors from the final household trip and person trip files were then added to the file. Two files were generated: NUW13CMP file which contained 6,614 household trip records and PRW13CMP file which contained 8,530 person trip records. Eight hundred seventy-nine records from the original W1C3 file could not be matched with the final trip files received from the Census. Among them, 559 of the 879 records have flags indicating out-of-scope (such as trips with one-way distance shorter than 100 miles, foreign origin trips, or marked for elimination due to problems that could not be solved within the quick-edit) so they were not expected to be included in Census's final files. No problems are evident in the remaining 320 trip records, thus the reason for their exclusion is unknown.

After eliminating the 879 non-matching records, the W1C3 study sets (NUW13CMP and PRW13CMP) contained trip records representing a total of over 14 million household trips and about 18 million person trips. These records correspond to approximately 25 percent and 22 percent of the total W1C3 household trips and person trips, respectively. ORNL would have been unable to complete distance calculations or calculations would have been incorrect without editing of these data.

If processed with AUTOFIX alone, approximately 402 thousand household trips, or 510 thousand person trips from W1C3 would have been deleted due to missing or insufficient location or mode information. This accounts for roughly three percent of the total trip records in each of the modified W1C3 files, NUW13CMP and PRW13CMP. Addition of the manual quick-fix processing also changed mode-distances for an additional 1 million household trips and 4 million person trips, beyond those fixed by the AUTOFIX process alone (see Tables 16 and 17). This is equivalent to about 18 percent for household trips and 21 percent for person trips that were edited under the revised processing method (i.e., files NUW13CMP and PRW13CMP, respectively).

As expected, the majority of the data quality improvements resulted from the systematic AUTOFIX procedure. The AUTOFIX process alone improved the quality of the data for about 19 percent of the household trips and about 17 percent of the person trips in the final W1C3 data set. "Manual quick-fix" procedures improved the quality of an additional 6 percent of the total W1C3 household trips and 5 percent of the total W1C3 person trips. Overall, the revised processing method contributed to an improvement of 25 percent for household trips and about 22 percent of the person trips, which accounted for a total of 53 billion person miles in W1C3.

Table 16. Edits on the W1C3 Study Set

	# household trips	% edited (household trips)	# person trips	% edited (person trips)
total in final W1C3	57,782	na	83,674	na
total in edited W1C3 by the revised method	14,415	24.9 %	18,362	21.9%
records either edited differently or not edited by AUTOFIX	3,282	5.7 %	4,168	5.0 %
number deleted in AUTOFIX but saved by the revised method	402	0.7 %	510	0.6 %

Table 17. Edits on the W1C3 Study Set

	trip-miles	% edited (trip-miles)	person-miles	% edited (person-miles)
total in final W1C3	63,118,973	100 %	86,639,836	100 %
total in edited W1C3 by the revised method	38,522,509	61.0 %	49,871,473	57.6 %
records either edited differently or not edited by AUTOFIX	9,791,866	15.5 %	12,759,901	14.7 %
number deleted in AUTOFIX but saved by the revised method	717,080	1.4%	882,467	1.0 %

7.4 RESULTS FROM THE COMPARISON STUDY

The study found that the AUTOFIX combined with the manual quick-fix processing provided a reasonable trade-off between minimizing the processing time and maintaining the quality of the data set. The AUTOFIX program could be useful for the next ATS, but detailed checking and editing on a limited set of ATS 2000 data (as was done for W1C1 of ATS 95) would be beneficial to assist in updating the databases for use with the new data. A lengthy detailed checking and editing process can be eliminated if an initial investment of time and resources is undertaken to assemble up-to-date databases for the AUTOFIX process.

8. RECOMMENDATIONS AND LESSONS LEARNED

This section describes recommendations and lessons learned by ORNL from experiences and knowledge gained during the processing of the 1995 ATS data. This information could be useful in working with future ATSs as well as other similar surveys. In fact, some of these recommendations and lessons learned were adopted during the processing of data from the 1997 CFS, which collected data on freight shipments from a sample of establishments engaged in mining, manufacturing, wholesale, and selected retail and service activities. Among the items of interest reported for each CFS shipment were the city, state, and zip code of the origin and destination; the mode(s) of transportation, and the weight and value of the shipment. ORNL's role in the 1997 CFS was to estimate shipment distance, which was not reported in the CFS.

8.1 GENERAL

Roughly one quarter of the household trip records, or equivalently one-fifth of the person trip records, in the 1995 ATS were modified by ORNL. These changes or modifications were required in order to improve data quality, perform distance calculations, or to prevent elimination of data. Many of these editing and imputation results were stored in databases or integrated in utility programs and used for subsequent Wave-Cycle data sets. In addition, many of the procedures, utility programs, and databases developed by ORNL during the 1995 ATS processing were also used during the processing of the CFS 1997 data.

During the processing of the 1997 CFS data, the ORNL project team used extensive knowledge gained from the 1995 ATS. They applied the ATS flow-based data processing procedures and used the "SSE" model to better utilize ORNL's transportation expertise. They also applied the edit log used in ATS to identify what record field was edited, when, and by whom. These procedures not only improved the data quality of CFS 97, but also reduced data processing time and, consequently, reduced costs. With the lessons learned from the CFS 93 and the ATS 95, the cost of distance calculation for CFS 97 was reduced by more than 50 percent of the cost for CFS 93. The following is a general summary of lessons learned from the processing of ATS data.

During the processing of ATS data, it was found that trips with Alaska as the origin or the destination required special effort for distance modeling. Many of the local roads were not in the national highway network. For these cases, distance was calculated using GCD and a circuity factor. Also, some of these trips used small local airports, which were not available in our database. These airports were identified, geocoded, and added to the airport database.

Trips with Hawaii as the origin or the destination often required special effort as well. City place names in Hawaii are confusing, and the misspellings of place names are common. ORNL's "ZIPFIXED"

database (used by the AUTOFIX utility program) contains many matches for these place names. This "ZIPFIXED" database was utilized during the processing of 97 CFS data and should be useful for geocoding tasks in future ATSs.

Furthermore, ORNL developed a foreign country place name database with lists of foreign city/ country matches and foreign city/country misspellings matches that could be used for the next ATS. The 1997 CFS project used the ATS foreign country database as a base in identifying port locations for export shipments.

Certain locations were frequently mistaken to be in a foreign country, for example, locations in Hawaii and New Mexico. In addition, places in California (CA) were sometimes coded as if they were in Canada. This type of error typically occurred because of mistakes in the 2-digit state code. Other common mistakes included confusion in the 2-digit state code for Alaska, Alabama, Arizona, and Arkansas as well as Mississippi, Michigan, Minnesota, and Missouri (e.g., mistakes such as "MI" for Mississippi, "MS" for Missouri, "AL" for Alaska, or "AK" for Arkansas). ORNL's "REALYUSA" and "ZIPFIXED" databases (used by the AUTOFIX program) contains many matches for these.

Respondents often reported place names for their destinations with names of state parks, casinos, lakes, etc. Therefore, associated zip codes were not coded by Autocoder or the follow-up clerical coding at the Census. ORNL's "ZIPFIXED" database contains many matches for these.

ATS data also showed that in many occasions respondents reported an incorrect mode of transportation as the main mode used during a trip. This was especially true for trips that involved cruises where the longest traveled segment was, in fact, by air or highway to the port of exit, yet the respondent reported water as the main mode. This inaccuracy might be significant if future surveys obtain only information about the main mode of a trip.

Many of the data problems identified during the ATS processing could be eliminated or fixed by the use of a computer-aided system with built-in "intelligence" for the interviewer. This type of system could provide pull-down selection windows for location information to help identify geographic locations. It could also display trip segments on a map as data are entered, and the feasibility of the reported mode of transportation could be checked at the same time.

8.2 TRIPS THAT NEEDED CLARIFICATION

The following provides a list of types of trips that seemed to cause the most confusion for the interviewer. Therefore, they might not have been handled consistently by interviewers. Questions associated with the collection of data for these types of trips need further clarification before the next ATS.

- Trips that included a cruise ship
  There needs to be a lodging type for cruise ships added to the survey. Also, the "farthest place traveled" questions are confusing to cruise ship passengers

- Trips to and from college
  There was confusion over whether to count this type of trip as a commute or a regular trip.

- Trips for National Guard duty
  There was confusion about whether or not these were "Active Duty" trips and, therefore, were not to be entered into the survey.

- Hiking trips
  Usually hikers are not close to towns or are unsure what town they are near when they stop for the night.

- There was confusion about trips where household members start out together but split up along the way.

- There was confusion about whether a trip was a commute trip if a person spent the night outside his/her home.

8.3 THE MEMO FIELD

Additional information that the respondent provided about a trip or comments about the survey that did not fit into pre-defined fields were captured in memo fields. These fields were sometimes called "F7" fields because the interviewer could press the F7 key on the computer at any time during the computer-aided telephone interview (CATI) or computer-aided personal interview (CAPI) to enter additional information. This information could come from either the respondent or the interviewer. For example, the interviewer could use memo fields to record that a respondent objected to a specific question(s), such as those about age, income, or children in the household. The respondent might provide additional information about a stop along the way or modify a previous response. This information would be captured in these memo fields. Interviewers also used memo fields to enter notes or comments of their own. Typical examples of interviewers' information captured here was that a respondent seemed reluctant to answer questions or that the survey instrument was not performing correctly (e.g., pulling up screens to capture information about a move when the respondent had not moved into the household).

Because memo fields were used to capture information from the respondent and the interviewer and because it captured so many different kinds of information, it appears that this information was not used to increase the accuracy of the survey responses. ORNL briefly evaluated these memo fields and summarized them into a "memo" database. Table 18 shows an analysis of these

Table 18. Responses from the Memo Field

Response Type	Examples	% of Records in Memo Database
Indicated missing trips	Forgot about a trip to visit mother	~4%
Provided information about stops	Stopped at a specific place to spend the night	~18%
Gave date corrections	February 18 rather than February 2	~2%
Gave more information on dates	The second stop was on March 13	~6%
Gave more information about number of nights	Spent 2 nights at the first stop; spent 6 nights at the destination	~11%
Refused to continue with the survey	Wants off the phone; survey is too long; information is business confidential	~3%
Provided additional information about education, HH income, employment, race, etc.	Received her masters degree this year.	< 1%
Gave additional information on mode	Took a bus tour to a specific place	~13%
Gave information on lodging	Spent 3 nights at the Holiday Inn there.	~8%
Gave reason for the stop	On this trip also stopped at a national historic park for scenic and educational reasons	~2%
Gave mode change information	Arrived by direct fight and charter bus took them to the steamboat dock	~1%
Gave information about commuter trips	Commutes twice daily; commuting trips are on weekends each month	~3%
Gave information on number of people on the trip	Three additional family members were on the trip	~8%
Complained about system problems	Could not make trip corrections or changes or add additional information	~8%
Indicated they were moving out	Will be moving to another house in 2 months	~1%

responses from 4,238 records, which included responses collected after the first Wave of Cycle 1 data (i.e., C1W1 memo field data was not provided to ORNL).

The information given in the memo field might have been useful, but time and budget constraints prevented its use in evaluating possible changes to the data collected for the 1995 survey.

8.4 CONCLUSIONS

Based on experiences and lessons learned from working with ATS 95 data, we developed and summarized in this section several recommendations for future ATSs, or similar surveys.

An add-on capability that enables the interviewer to catch inaccuracies or mistakes in location or mode information while the interview is on-going would be highly beneficial to the survey instrument. This capability would significantly help to reduce response or typing errors and therefore improve data quality. Time and costs associated with editing and fixing data errors would be dramatically reduced.

The AUTOFIX program, or a similar utility program, could be useful for the next ATS as well. Detailed checking and editing on a limited set of ATS 2000 data (as was done for W1C1 of ATS 95) would be beneficial to assist in updating the databases for use with the new data. A lengthy detailed checking and editing process can be eliminated if an initial investment of time and resources is undertaken to assemble up-to-date databases for the automatic corrections.

Some level of a "manual quick-fix" of the data should be performed to gather information necessary to enhance and update databases or program codes used in the automatic checking/editing procedures (e.g., AUTOFIX). This will allow a large number of trip records in subsequent sets to be edited and recovered by the updated automated processing. To prevent the loss of trip records with relatively minor problems, a manual quick fix should also be used to handle obvious, but case-specific, errors that can not be fixed with an automated process.

Possible future studies could include a more detailed analysis of the information contained in the ATS 95 memo fields. This type of analysis could help to identify survey questions that caused the most respondent confusion or those that needed further refinement.

With the experience gained from this project and by using the tools and databases developed, the team at ORNL feel that a dramatic reduction in time and cost for processing future ATS data can be expected. We estimate that this reduction in cost can save at least 50 percent as compared to 1995 ATS processing.

9. REFERENCES

5-Digit Zip Code File, Version 3.2, a product of the Geographic Data Technology, (GDT), Inc.

ATS Data Processing Procedures. (1996). Draft report prepared for the Bureau of Transportation Statistics and Bureau of Census, prepared by Oak Ridge National Laboratory, Oak Ridge, Tennessee, March, unpublished.

Bronzini, M.S., Chin, S-M, Liu, C., Middendorf, D.P., and Peterson, B.E. (1996). *Methodology for Estimating Freight Shipment Distances For The 1993 Commodity Flow Survey*. Report prepared for the Bureau of Transportation Statistics and Bureau of Census, by Oak Ridge National Laboratory, Oak Ridge, TN 37831, April.

Bureau of the Census, Public Law (P.L.) 94-171 Data on CD-ROM (PL-94),.

Digital Chart of the World, Edition 1. A digital database of the world developed by the United States Defense Mapping Agency (DMA) with the cooperation of Australia, Canada, and the United Kingdom.

Geographic Names Information System (GNIS), US Geographic Services, *http://www-nmd.usgs.gov/www/gnis/*

Impact of ORNL Quality Assurance on the 1995 ATS Data Trip. (1997). Draft report prepared for the Bureau of Transportation Statistics and Bureau of Census, prepared by Oak Ridge National Laboratory, Oak Ridge, Tennessee, December, unpublished.

Kendall, M. G., and Moran, P. A. P. (1963). *Geometric Probability*, Griffins Statistical Monographs and Course No. 10, ed. Kendall, London.

MacMillan World Atlas, The. (1996). A Simon and Schuster MacMillan Company, New York, NY.

Southworth, F., Peterson, B.E, and Chin, S-M. (1998). *Methodology for Estimating Freight Shipment Distances For The 1997 Commodity Flow Survey*. Report prepared for the Bureau of Transportation Statistics and Bureau of Census, by Oak Ridge National Laboratory, Oak Ridge TN 37831, October.

Suggestions for ATS 2000. (2000). Technical memorandum prepared for the Bureau of Transportation Statistics and Bureau of Census, prepared by Oak Ridge National Laboratory, Oak Ridge, Tennessee, January, unpublished.

Times Atlas of the World, The, *Ninth Comprehensive Edition*. (1994). Times Books, a Division of Random house, Inc. New York and Canada.

APPENDIX A: EDIT VARIABLES (ORNL Internal Use)

New variables created during the data editing process of ATS data are listed below. All new variables are appended to the existing record so that original data was kept and unchanged.

New ORNL Variable	Census Variable	Code Definition or Variable Description
.		
DLTFLAG	--------	"1" - trip deleted due to the lack of origin information "2" - stop deleted due to insufficient information "3" - origin was in a foreign country (record Type 1, 9, and 10) "4" - trip deleted when noted as unwanted record in the file "5" - stop record deleted; same O-D locations "9" - one-way calculated distance for the trip was <100 miles
CHANGED	--------	"1" - one or more changes made to the segment record Blank - no changes to the segment record

ODFLAG	--------	This field consists of a 2-digit character code. The 1st digit is for the origin information. The 2nd digit is for the destination information. The following codes are used to identify any editing/imputation made to either the origin or the destination data field:
		    "0" - no changes
		    "1" - only state was known (imputed zip code)
		    "2" - only city was known (imputed zip code)
		    "3" - city and state both unknown (imputed zip code)
		    "4" - edited zip code for a specified city and state but unknown (city /state not found) zip code
		    "5" - for <u>foreign country cases</u> where city unknown; imputed location (usually Canada and Mexico)
		For example: ODFLAG = "20" representing origin zip code was imputed due to lack of state information but the destination information was unchanged; ODFLAG = "13" means the origin zip code was imputed due to lack of city name and the destination zip code was also imputed because its location was not provided.
		
		A code should be entered for this field if any of the following fields has non-blank entry: NEWOCITY, NEWOST, NEWOZIP, NEWDCITY, NEWDST, NEWDZIP.

NEWOCITY	OCITY	Origin city trip/segment
NEWOST	OST	Origin state

NEWOZIP	OZIP	Origin zip
NEWDCITY	DCITY	Destination city
NEWDST	DST	Destination state
NEWDZIP	DZIP	Destination zip
ODFCFLAG	--------	This field is a 2-digit character code for origin and destination foreign country information . The 1st digit is for the origin foreign country information. The 2nd digit is for the destination foreign country information. The following codes should be used for each O-D pair: "0" - no changes "1" - foreign country name was edited "2" - missing foreign country name imputed by ORNL A code should be entered for this field if either the NEWOFC or the NEWDFC field has non-blank value.
NEWOFC	FCTRYO	Origin foreign country
NEWDFC	FCTRYD	Destination foreign country
NEWRTYPE	RECTYPE	Record type
NEWSTOPN	STOPNUM	Stop number
NEWMODE	--------	Edited mode of transportation for the given segment record. This data and its associated record type determines values for the following variables: TRNS1, TRS2, WTYP1, and WTYP3.
MODEFLAG	--------	Flag field for the NEWMODE variable: "1" - edited "2" - imputed
NEWNITE	--------	Total number of nights spent at destination/stop. This data and its associated record type determines values for the following fields: CNITE, STP1N, and STP3N.

NITEFLAG   --------   Flag field for the NEWNITE variable:
                                                                  "1" - edited
                                                                  "2" - imputed

APPENDIX B: VARIABLES INCLUDED IN THE OUTPUT FILES TO THE CENSUS

Fields needed to identify a trip

      CASEID   case identification number to identify a household in a given Cycle

      CNTRLNUM   household control number

      PERSONS   person number used to identify an individual in a household

      MDEST   trip number used to identify trips taken by a given person in a household

Fields needed to identify a segment

      ORECTYPE   record type, from the original Census's trip file

            1- total going trip (i.e., home to final destination)
            2- home to departing station
            3- departing station to final station
            4- stop segments for the going trip
            5- side trip segments
            6- stop segments for the returning trip
            7- total return trip (i.e., final destination to home)
            9- commuter trips
            10- mover trip from old home to new home
            11- mover trip, home to departing station
            12- mover trip, station to new home
            13- mover trip, stop segments

      RECTYPE   record type, updated by ORNL
                                  (Same code as above)

      OSTOPNUM   the original 2-digit stop sequence number from the trip file

      STOPNUM   stop number updated by ORNL. If no new stop number was added and no resequencing of stops was done to a given trip record, the value in this field should equal the value in OSTOPNUM multiplied by 10. That is, the updated stop number uses a 3-digit integer to identify segment sequence within a record type for a given trip. The first two digits in the new stop number represent a sequence number the same as those used in the original stop number (1, 2, 3, ..., 28, 29, 30, ...). The third digit serves as a "subscript" to identify the two sub-segments split by either

border crossing or mode changes. Under this format, if the STOPNUM ends in "0", it indicates no split on the given segment was done. For example, a trip with 2 stops will have STOPNUM codes as "10", "20", & "30" for the 3 segment records. However, suppose the second segment crosses the U.S. border; it will need to be split into domestic and international portions for distance calculation. The final trip file, therefore, has 4 segment records with STOPNUM coded as follows: "10", "21", "22", & "30".

Trip/segment deletion flag

DLTFLAG record marked for deletion:

  1 - trip deleted due to lack of original information
  2 - segment deleted due to insufficient stop information
  3 - trip deleted due to foreign country origin
  4 - trip deleted when noted as "unwanted" record in the data file
  5 - segment deleted due to replicated O-D in segment record
  9 - trip deleted due to one-way travel distance < 100 miles

Fields with O-D information

OCITY origin city name

OST origin 2-digit state name

OZIP origin 5-digit zip code
  Note: zip code field will be blank for
    - foreign country location
    - station/airport/terminal
    - border crossing location
    - the place where split of a segment was required due to a mode change
    when a respondent did not report such place as a stop

OFLAG origin flag, 1 = edited; 2 = imputed; & blank if no change

OFC_CODE origin 3-digit foreign country code

OFCFLAG origin foreign country flag, 1 = edited; 2 = imputed; & blank if no change

DCITY destination city name

DST destination 2-digit state name

88

DZIP   destination 5-digit zip code
                (see notes in OZIP)

DFLAG   destination flag, 1 = edited; 2 = imputed; & blank if no change

DFC_CODE   destination 3-digit foreign country code

DFCFLAG   destination foreign country flag, 1 = edited; 2 = imputed; & blank if no change

Fields contain mode of transportation information

SEG_MODE   mode of transportation used for the trip segments other than Type 2:

                H - highway mode (01, 02, 03, 10, 15, 16, 17, 18)
                B - highway mode, bus (06, 07, 08)
                A - air mode, commercial (04)
                P - air mode, personal/corporate (05)
                T - train/Amtrak (09)
                W - waterway mode (11, 12, 13, 14)

                All Type 2 records are assumed to the highway mode, i.e., codes 'H' for
                SEG_MODE.

MODECODE   mode of transportation used for the segment, '01' - '18' for record types
                    other than 2; '1'-'10' for record Type 2

MODEFLAG   flag associated with the variable MODECODE, 1 = edited; 2 = imputed;
                    & blank if no change

Fields relate to trip/segment distances

GCD_OD   Great circle distance from origin to destination for the given <u>record</u>

EST_US   for all record types other than 1, 7, or 10:
                    Calculated distance from segment origin to segment destination - U.S.
                    portion; 0 if international segment

                    for record Type 1, sum of domestic Type 4 records
                    for record Type 7, sum of domestic Type 6 records
                    for record Type 10, sum of domestic Type 13 records
                    (sum of all U.S. segment distances for the given direction of a trip)

EST_INTL	for all record types other than 1, 7, or 10:
	calculated distance from segment origin to segment destination - international portion; 0 if U.S. segment
	for record Type 1, 7, and 10:
	sum of all international segment distances for the given direction of a trip; 0 if all segments are U.S.
GCD_US	for record Type 1, 7, & 10:
	sum of all U.S. segment GCDs
	all other record types:
	the same as in GCD_OD if U.S. segment, 0 if international segment
GCD_INTL	for record Type 1, 7, & 10:
	sum of all international segment GCDs
	all other record types:
	the same as in GCD_OD if international segment; 0 if U.S. segment

Additional information for the segment record

NEWNITE	new value for the number of nights spent at destination of the segment; blanks if original value was not modified
NITEFLAG	flag for the NEWNITE, 1 = edited; 2 = imputed; & blank if no change
NEWRDMS	new value for the road-missed nights when traveling overnight on the segments; blanks if original value was not modified
RDMSFLAG	flag for the NEWRDMS, 1 = edited; 2 = imputed; & blank if no change
SEQCHNG	reason for changes on stop segments

	blank -	no change
	1 -	add-stop only (original Type 8 add-stop was inserted to the trip without resequence of stop segments
	3 -	reorder out-of-sequence stops only (no new stop was added)
	5 -	both Add Stop and resequence of stop segments

| WHYSPLIT | reason for splitting a given segment (i.e., a non-zero at the third-digit position in the variable STOPNUM |

blank -   no change

14 -   border crossing, going trip
15 -   border crossing, side trip
16 -   border crossing, returning trip
22 -   station added for access segment
23 -   imputed Type 3 record to separate multiple modes used
                in a given segment during the going trip
24 -   imputed Type 4 record to separate multiple modes used
                in a given segment during the going trip
26 -   imputed Type 6 record to separate multiple modes used
                in a given segment during the returning trip
31 -   mover trip, station added for access segment (i.e.,
                imputed Type 11
32 -   imputed Type 12 record to separate multiple modes used
                in a given segment during the moving trip
33 -   imputed Type 13 record to separate multiple modes used
                in a given segment during the moving trip

APPENDIX C: DATA REQUIREMENT FOR THE "PATH" FILE (ORNL Internal Use)

In order to produce needed travel flows for the ATS project, "path" found for each trip segment by the distance calculation programs will be stored in files. The following is a list of data items required in this "path" file:

> Case ID
> Person number
> Trip number
> Record type
> Stop number
> Mode of transportation for the segment
> Segment beginning node number
> Segment ending node number
> List of link numbers (link_1, link_2, ...)

No international segment path will be required.

Note: Due to the extremely large disk storage space required by the path file, this process was performed only on an as-needed basis.

INTERNAL DISTRIBUTION

1.	G.E. Courville	11–15.	J.P. Rollow
2.	T.R. Curlee	16.	R.B. Shelton
3.	D.L. Greene	17.	Central Research Library
4.	P.S. Hu	18.	Document Reference Section
5–9.	H.L. Hwang	19.	Laboratory Records
10.	C.I. Moser	20.	Laboratory Records—RC

EXTERNAL DISTRIBUTION

21. L.A. Abron, President, PEER Consultants, P.C., 1460 Gulf Blvd., 11th Floor, Clearwater, FL 34630

22. S.L. Cutter, Director, Hazards Research Lab, Department of Geography, University of South Carolina, Columbia, SC 29208

23. S.G. Hildebrand, Director, Environmental Sciences Division, Oak Ridge National Laboratory, Post Office Box 2008, Oak Ridge, TN 37831-6037

24. P.R. Rittelmann, FAIA, Executive Vice President, Burt Hill Kosar Rittelmann Associates, 400 Morgan Center, Butler, PA 16001-5977

25. S.F. Tierney, The Economic Resource Group, Inc., One Mifflin Place, Cambridge, MA 02138

26. C.M. Walton, Ernest H.Cockrell Centennial Chair in Engineering and Chairman, Department of Civil Engineering, University of Texas at Austin, E. Cockrell, Jr. Hall I, Suite 4210, Austin, TX 78712-1075

27–36. OSTI, U.S. Department of Energy, P.O. Box 62, Oak Ridge, TN 37831