Final Report

Analysis of GPS-Based Data for Light Duty Vehicles

Contract No. 04-327 UCR

August 2006
Revised: January 2007

Prepared for:

Hector Maldonado
Research Division
California Air Resources Board
1001 I Street, 5th Floor
Sacramento, CA 95814
hmaldona@arb.ca.gov
(916) 445-6015

Principal Investigator Matthew Barth
Contributing Authors: Weihua Zhu, Kanok Boriboonsomsin, Luis Ordonez
College of Engineering-Center for Environmental Research and Technology
University of California
Riverside, CA 92521
(951) 781-5782
(951) 781-5790 fax

Disclaimer

Acknowledgments

The authors thank the following organizations and individuals for their valuable contributions to this project. First and foremost, we would like to acknowledge the California Department of Transportation (CALTRANS) for providing GPS-based vehicle activity data from their 2001-2002 California Statewide Household Travel Survey Study. Similarly, we would like to acknowledge the Southern California Association of Governments (SCAG) for sharing their Year 2000 Post-Census Regional Travel Survey for Southern California. These datasets formed the basis for much of this work. In addition, other GPS-based vehicle activity dataset have been assembled in this project from various sources, too numerous to mention.

We acknowledge the funding from the California Air Resources Board (CARB) to carry out this project. We acknowledge the technical assistance and support from the CARB staff involved in this study including Hector Maldonado, Jeff Long, and Steven Magbuhat.

Table of Contents

Executive Summary

Understanding vehicle activity patterns both spatially and temporally is critical for building accurate mobile source emissions inventories. Vehicle activity has frequently been characterized using average speed and vehicle miles traveled (VMT), however advances in modeling of mobile sources have increased the resolution in vehicle activity necessary for using the new models to their full capabilities. Recently, two GPS (Global Positioning System)-based vehicle activity datasets have become available from several different research programs. In 2001, the California Department of Transportation (Caltrans) conducted their 2001 California Statewide Household Travel Survey Program, which contains GPS-based data sets from across the state. This data set is approximately 125 MB in size and represents about 272 households. This database was divided into two parts, one corresponding to Northern California trips and the other corresponding to Southern California trips. Further, the Southern California Association of Governments (SCAG) carried out a post-census travel survey in 2001, which contains a number of GPS-datalogger datasets. This data set represents approximately 467 households.

The University of California College of Engineering-Center for Environmental Research and Technology (CE-CERT) has had extensive experience in designing, conducting, and performing data analysis for vehicle activity studies. In this research project, the CALTRANS and SCAG datasets were extensively examined, extracting vehicle activity information for better characterization of vehicle activity. These datasets were originally created to provide backup information to written travel surveys. Since these data acquisition programs were carried out across a wide range of representative households, it was felt that under further analysis, these data could provide useful information in characterizing vehicle activity for creating accurate mobile source emissions inventories. Three primary research tasks were carried out: 1) Specific vehicle activity data sets were acquired, pre-processed, and organized into a vehicle activity database; 2) Data analysis was performed on the two vehicle activity datasets, consisting of an examination of vehicle starts, as well as global trip characteristics (e.g., number of trips, trip length, and diurnal patterns; for comparisons between Northern and Southern California, the Caltrans database was utilized exclusively); and 3) A detailed speed-acceleration histogram analysis was performed, examining the velocity trajectory characteristics in the vehicle activity database. This was performed across different roadway facility types.

Based on this analysis, several general conclusions were made: 1) The average distance per trip was relatively short (4 to 6 miles) with an average trip duration around 8 to 12 minutes, with trips slightly shorter in distance but longer in duration in Southern California compared to Northern California (based on the Caltrans dataset); 2) The number of trips per day per vehicle was approximately 5 for both data sets; 3) The household-based datasets showed that there were little differences of travel from Monday – Friday, however on Saturdays and Sundays, the trips were significantly reduced—this is likely due to the fact that very little data was collected over the weekends; 4) An analysis of the diurnal trip patterns for the two household datasets did not show a typical commute pattern with a distinctive AM morning peak and a PM afternoon peak. Instead,

most activity peaked during the early afternoon in a single mode distribution; 5) An analysis of the soak time periods of the vehicles showed a two-mode distribution, where one peak occurring for 10 minutes or less (30% of the distribution and the other less pronounced peak occurring in the range of 120 – 360 minutes (13.5% of distribution); 6) After disaggregating the dataset by roadway facility type, it was seen that approximately 55% - 65% of VMT occurs on freeways, and the remaining 35% -45% occurs on surface streets; 7) In contrast, trip time spent on highways is approximately 35% - 45% while for surface streets, it was approximately 55% - 65%; 8) Average speeds were significantly higher on highways (as expected) compared to surface streets—Northern California had slightly higher speeds overall (based on the Caltrans dataset); 9) A number of speed-acceleration parameters and speed-acceleration frequency distributions were evaluated across the vehicle activity databases; as expected, surface streets displayed greater speed-acceleration fluctuation compared to highway travel.

Based on this study, some of the results can be used to update the vehicle activity portion of CARB's emissions inventory process. Several additional recommendations are as follows:

1) It is now possible to create roadway facility specific emission factors for the different kinds of driving that occur on the different road types. This can be accomplished by taking the driving snippets from the corresponding facilities and running them through a modal emissions model (weighted for a specific fleet). This would then allow for a link-based emissions inventory process where activity is measured on a link-by-link basis then multiplied by the corresponding emissions factor.

2) Similarly, it is possible to create representative "driving cycles" that correspond to specific roadway facility types. These driving cycles could also be used to create facility-specific emission factors through a real-world test program.

3) Now that appropriate analysis tools have been developed for processing GPS-based vehicle activity datasets, it is now possible to carryout additional vehicle activity studies at a fairly low cost. It is not very expensive to put in GPS dataloggers into representative vehicles and use those vehicles as "probes" to determine traffic and activity conditions. Of particular interest would be a truck travel pattern study.

4) Hybrid electric and plug-in hybrid electric vehicle energy management strategies can be optimized using these vehicle activity data. Since the vehicle activity is representative of real-world driving patterns, the energy management of a charge-sustaining strategy or zero emission range can be optimized based on the data sets.

1. Introduction and Background

There are three critical components that are used in developing accurate mobile source emissions inventories: 1) vehicle emission rates; 2) vehicle activity; and 3) vehicle fleet distribution. To date, much effort has been spent on developing accurate real-world vehicle emission rates, however less effort has been applied to correctly identify real-world vehicle activity patterns and vehicle fleet distributions. Vehicle activity has frequently been characterized using average speed and vehicle miles traveled (VMT), however advances in modeling of mobile sources have increased the resolution in vehicle activity necessary for using the new models to their full capabilities. In order to improve our understanding of vehicle activity patterns, various research studies have taken place in recent years. For example:

- In the mid-1990s, both the U.S. EPA and CARB recognized that existing certification driving cycles (e.g., the Federal Test Procedure) were not very representative of modern traffic driving patterns, leading to the development of improved driving cycles such as CARB's Unified Cycle [Gammariello & Long, 1996] and U.S. EPA's SFTP (supplemental federal test procedure, see [U.S. EPA, 1993]).

- A new generation of research emission inventory models is now being developed and applied to a variety of cities, including many cities outside the U.S. These models, such as Georgia Tech's MEASURE model [Guensler et al., 1998], the U.S. EPA MOVES model [Hart et al., 2002; Koupal et al., 2002], UC Riverside's CMEM [Barth et al., 1999; Malcolm et al., 2002] and IVE (International Vehicle Emissions) model [Davis et al., 2005], rely not on certification cycles for emissions and activity estimates, but rather on real-world driving patterns.

- In order to improve the spatial and temporal qualities of CARB's EMFAC model, there have been specific studies aimed at examining differences between typical weekday and weekend travel (see, e.g., [STI, 2004]).

In many of these programs, Global Positioning System (GPS) datalogging units were often used to capture vehicle activity patterns of the vehicles. These dataloggers are installed in individual vehicles and are capable of recording position, speed, and the time of a vehicle's activity. These are extremely useful tools for collecting in-use driving pattern data. From these data, other information useful for emissions modeling can be derived, such as [Malcolm et al., 2002]:

- number of vehicle starts;

- time/location of vehicle start and stop events;

- trip length statistics (e.g., VMT per trip, VMT per day); and

- typical velocity trajectory characteristics (e.g., average, maximum, and instantaneous speeds).

In this project, two relatively new vehicle activity datasets have been examined, and a variety of information has been extracted. These vehicle activity datasets are described in Table 1.

Source	Description
CALTRANS	In 2001, the California Department of Transportation (Caltrans) conducted their 2001 California Statewide Household Travel Survey Program, which contains GPS-based data sets from across the state. This data set is approximately 125 MB in size and represents about 272 households.
SCAG	The Southern California Association of Governments (SCAG) carried out a post-census travel survey in 2001, which contains a number of GPS-datalogger datasets. This data set represents approximately 467 households. [NuStats, 2002]

Table 1.1: Description and Source of Vehicle Activity Datasets

The University of California College of Engineering-Center for Environmental Research and Technology (CE-CERT) has extensive experience in designing, conducting, and analyzing GPS-based vehicle activity data studies. For example, in 2001, over 13 million seconds of GPS-based vehicle trajectory data were collected in the Los Angeles region to better understand the impacts of low emitting vehicles [Malcolm et al., 2002]. Further, CE-CERT has conducted vehicle activity projects in India, China, Africa, and South America as part of their international emissions modeling efforts [Davis et al., 2005]. CE-CERT has utilized their developed expertise in carrying out this research project.

In this project, three primary tasks were carried out:

1) Specific vehicle activity data sets were acquired, pre-processed, and organized into a vehicle activity database;

2) Data analysis was performed on the two vehicle activity datasets listed in Table 1.1. This data analysis consisted of an examination of vehicle starts, as well as global trip characteristics (number of trips, trip length, diurnal patterns, etc.);

3) A detailed speed-acceleration histogram analysis was performed, examining the velocity trajectory characteristics in the vehicle activity database. This was performed across different roadway facility types.

Chapter 2 of this report briefly describes the developed vehicle activity database and the analysis techniques employed. Chapter 3 describes the data analysis results. Chapter 4 provides further results of the velocity trajectory characteristics across different roadway facility types. Chapter 5 provides brief conclusions and recommendations for future work.

2. Database Design and Procedures

As described in Chapter 1, greater importance is being placed on vehicle activity datasets in order to improve mobile-source emissions inventory modeling. With the advent of GPS dataloggers, it is possible to collect a large amount of vehicle activity data from a wide variety of vehicles. Numerous vehicle activity studies have been carried out in the last several years, however there really hasn't been a consistent database design for vehicle activity. As part of this project, we have developed a three-level database structure, described below. With this design, the raw GPS-based vehicle trajectories (i.e., position and speed versus time) is subsequently processed and stored in the database for further analysis. The procedures for filtering, processing, and analysis are described briefly below.

2.1. Database Design and Data Conversion

The initial task of this research project was to develop a common database format for vehicle activity data, followed by the conversion of different vehicle activity datasets into this format. A vehicle activity database format was designed and is illustrated in Figure 2.1. This database has a three-level structure, characterizing the different kinds of datasets.

The top level of the database is simply a listing of the ***program*** or ***project*** that took place to collect the vehicle activity data. It contains fields such as the name of the program, the dates of the program, how many vehicles were tested, and the type of vehicle activity parameters that were collected (e.g., speed, location, etc.). Each entry in this level has a pointer to the second level of the database.

The second level of the database corresponds to the ***data runs*** or ***trips*** that took place within the overall program. For example, a vehicle going out to collect activity data would correspond to a single trip entry in the second layer of the database. For each trip or run, various parameters are listed including date, start and end times, vehicle type, and data frequency. Each entry in this level has a pointer to the third level of the database.

The lowest (third) level in the database is the ***time series data*** itself. Each time series entry is pointed to from the second level. The time series data contains the time sequence of position (e.g., latitude and longitude), speed, altitude, as well as other parameters that might have been collected (e.g., engine parameters, GPS-quality, external measurements). Further, in this layer of the database, additional parameters may be derived based on further processing, such as a determination of roadway type, lane type, or even congestion level of the roadway.

Dataset title	Date	Number of vehicles	Number of seconds	Acquisition frequency	Parameters
Selev	142 days: 2001/2002/2003/2004	25	1,603,910	Mostly 5Hz with some 1 Hz	Speed, lat.; long.; alt.
......	......	……	……	……	……

Run	Data	Start time	Vehicle	Duration	Frequency
1	6/9/2004	1:03:54	motorcycle	14,105	1Hz
2	6/8/2004	21:52:17	bus1	15,655	1Hz
......	......	……	……	……	……

Date	UTC time	Latitude	Longitude	RTCM_D	Alt	Speed	……
6/9/2004	1:03:54	31.17111	121.4259	0	-18	0.124	……
6/9/2004	1:03:55	31.17135	121.4257	0	22.3	0.062	……
......	......	……	……	……	……	……	……

Figure 2.1. Database Topology

2.2. Data Preprocessing

Vehicle trajectories are typically collected using some type of GPS datalogger, as illustrated in Figure 2.2. The datalogger is placed in the vehicle with a GPS antenna attached to the roof of the vehicle so that it has clear access to the GPS satellites. Once acquiring a positional lock with the satellites, the GPS dataloggers record a GPS navigational stream of data, typically at 1Hz. These data streams are usually recorded as an ASCII text file that is later imported into a spreadsheet or database file.

Depending on the GPS datalogger, several conversion and filtering steps are often necessary. For example, preprocessing may include: (1) conversion of latitude and longitude into decimal degrees, (2) conversion of speeds from knots to mph or metric units, and (3) conversion of date and time from UTC to local date and time. In addition to the vehicle position and velocity, the dataloggers may also record the number of satellites in view and the HDOP (Horizontal Dilution of Precision, see [Farrell & Barth, 1999]) values for each record. In this case, the scope of the preprocessing task can be extended to flag and investigate invalid and suspicious data points. For example, position/velocity records may be deemed invalid if the number of satellites is less than three (a minimum of four satellites is required for a 3-D fix) or if the HDOP value is greater than five. The overall filtering process is illustrated in Figure 2.3.

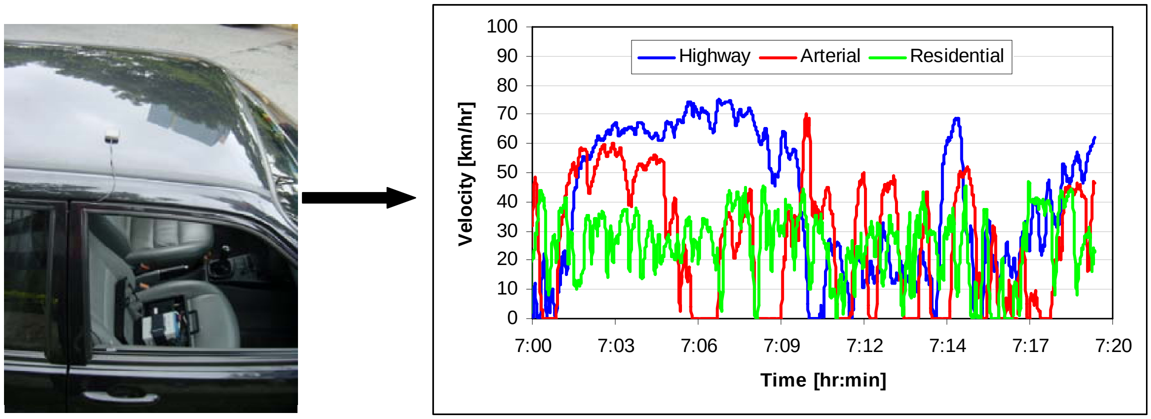

Figure 2.2. GPS vehicle trajectory data collection methodology: dataloggers are placed in vehicles and resulting data can then be analyzed across a number of parameters, such as speed patterns.

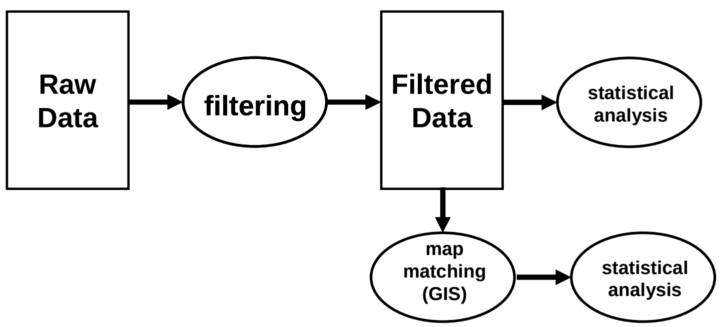

Figure 2.3. For this study, raw GPS trajectory data first undergo filtering to get rid of invalid points, followed by data filling or smoothing. The data is then examined to determine trip characteristics. Additional spatial processing is possible using map matching techniques in a Geographic Information System (GIS) environment.

2.3. Trip Detection and Characterization

After the GPS data streams are processed into a usable form, the next steps are to determine trip and trip characteristics. Generally, the GPS data streams are contiguous points in time. When there is a significant discontinuity in the time series, then that generally constitutes the end of an old trip, and the start of a new trip. Along with this time series discontinuity, velocity information is also examined. If the discontinuity occurred with a vehicle velocity of zero, then this is typically deemed a trip marker. If, on the other hand, the velocity of the vehicle was non-zero at the end and start of a (relatively short) time series discontinuity, then this usually indicates a loss of GPS data (such as going through a tunnel). One of the important parameters is a threshold that determines the duration for a time series discontinuity. This is often called the "dwell-time threshold" and a common threshold is 120 seconds [Bhat et al., 2005].

After detecting "trips" within the data stream, trip characterization must take place. This includes extracting trip-origin and trip-end locations, trip timing, trip distances, and trip speeds. Origin and destination trip-end locations may be determined by reading the location information from the first and last records of the GPS navigational stream corresponding to the trip. This can get a little tricky if GPS satellite acquisition is not immediate when a trip is first initiated (e.g., starting in a parking garage). However, the general rule of thumb is that the trip origin of one trip is typically the trip destination of the previous trip.

The trip start time is determined based on when the GPS device acquires its first fix after the start of the trip (i.e., from the time stamp on the first valid record for the trip). Similarly, the vehicle trip-end time is the time stamp on the last valid position assumed to be the end of the trip. Again, this can get tricky if there are some delays associated with GPS satellite acquisition.

For trip distance determination, there are two main methods that can be applied. The primary methodology is to calculate the point-to-point sum of each individual recorded position. This is fairly intuitive, just summing the distances between all points along the trip. However, it has been found in other studies that trip distance over-estimation sometimes occurs due to the positional inaccuracy of the GPS readings [Bhat et al., 2005]. A second method is to calculated distance based on the link distances in the roadway network data, after the GPS points have been matched to network links. This requires that the GPS traces be matched to an underlying road network to identify the actual links traveled by the vehicle. The trip length is determined as the sum of the length of all the roadway links traveled. This method can accommodate loss of signals along the trip much better, however it relies on successful map matching procedures, which is sometimes difficult in high-density roadway environments.

In terms of vehicle speed, nearly all GPS receivers record vehicle velocity based on Doppler measurements of the satellite signals. This is independent of the positional measurements and is general accurate to within 1 km/h. Average trip speed and measures of variations in speed along the trip length can be determined in a straightforward manner from these instantaneous speed measurements.

2.4. Trip Route and Roadway Facility Type Determination

As mentioned briefly above, map matching techniques (matching the GPS data points to appropriate links on an underlying GIS roadway network map) are generally required to perform trip route determination. This is also necessary for roadway facility type determination. It is important to note that map matching is not trivial, since both the GPS data and the digital roadway-network data have different levels of spatial accuracy and inherent errors. Consequently, the development of map-matching algorithms is in itself a very vast and complex field of study. Researchers have developed a wide array of methods using deterministic, probabilistic, and fuzzy-logic-based approaches (see for example, [TRB NCHRP Synthesis 301, 2001]) for matching GPS traces to GIS maps.

For the analysis carried out in this project, we utilized a very straightforward method of map matching, using ArcGIS 9 software from ESRI. This was used to determine roadway facility type of the data, as illustrated in Figure 2.4. In this figure, the latitude and longitude positional measurements are used by the map matching routine to determine a new (derived) parameter in the overall vehicle trajectory database.

Map matching

House_ID	Vehicle	GPS_ID	Latitude	Longitude	Speed	Time	Flag	Chain_SeqNo	Trip_SeqN	RoadType
2009745	1	1	32.78063	-115.5744	3.33	10/1/2001 6:51:19 PM	5	1	1	F
2009745	1	2	32.78062	-115.5743	7.87	10/1/2001 6:51:22 PM	0	1	1	F
…	…	…	…	…	…	…	…	…	…	…
2009745	1	4	32.78049	-115.5743	16.7	10/1/2001 6:51:24 PM	0	1	1	R
2009745	1	5	32.7804	-115.5743	19.54	10/1/2001 6:51:25 PM	0	1	1	R
…	…	…	…	…	…	…	…	…	…	…
2009745	1	7	32.78023	-115.5743	23.35	10/1/2001 6:51:27 PM	0	1	1	A
2009745	1	8	32.78014	-115.5743	24.92	10/1/2001 6:51:28 PM	0	1	1	A

Figure 2.4. Map matching process for creating a roadtype parameter in the time series dataset.

The main approach is to identify a group of vehicle activity data points that are spatially located within a specified lateral distance from the centerline of roads for a particular type of roadway. These two-staged procedures include: 1) data preparation and 2) data analysis. Each procedure is described in detail below.

As part of the data preparation, the standard TIGER/Line® 2000 roadway network dataset was used [ESRI, 2006]. The census 2000 TIGER/Line® database was converted to ESRI shapefile format for each county in California. Among several similar GIS roadway data sets, this set of data is most suitable for the purpose of our analysis because it includes the information of Census Feature Class Codes (CFCC) as one attribute in its database. CFCC designates hierarchical class codes to a group of defined features (e.g. roads, railroads, landmarks, etc). For instance, roads are designated A1 to A7 depending on their functional type. For our analysis, the roads layer for every county in Southern California was downloaded and aggregated to produce the complete roadway network for the region, as shown in Figure 2.5 (left). Then, each road type from A1 to A7 was extracted as a separate layer as shown in Figure 2.5 (right) for use in the data analysis.

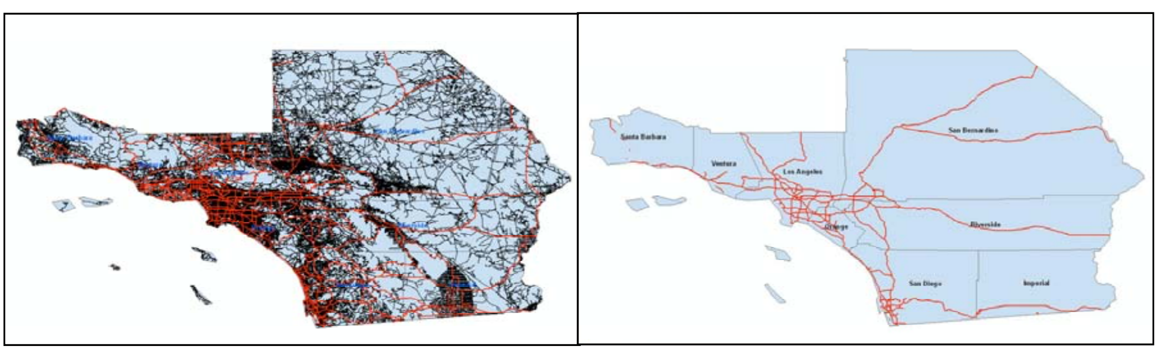

Figure 2.5. (Left) roadway network in Southern California; (Right) feature A1 (primary highway with limited access, interstate highways, and some toll highways)

The general characteristics of the A1 – A7 roadway types are provided in Table 2.1.

Road Class	Description
A1	Primary Highway With Limited Access Interstate highways and some toll highways are in this category (A1) and are distinguished by the presence of interchanges.
A2	Primary Road Without Limited Access Includes nationally and regionally important highways that do not have limited access as required by category A1. It consists mainly of US highways, but may include some state highways and county highways that connect cities and larger towns. A road in this category must be hard-surface (concrete or asphalt). It has intersections with other roads, may be divided or undivided, and have multi-lane or single-lane characteristics
A3	Secondary and Connecting Road Includes mostly state highways, but may include some county highways that connect smaller towns, subdivisions, and neighborhoods. The roads in this category generally are smaller than roads in Category A2, must be hard- surface (concrete or asphalt), and are usually undivided with single-lane characteristics. These roads usually have a local name along with a route number and intersect with many other roads and driveways
A4	Local, Neighborhood, and Rural Road A road in this category is used for local traffic and usually has a single lane of traffic in each direction. In an urban area, this is a neighborhood road and street that is not a thoroughfare belonging in categories A2 or A3. In a rural area, this is a short-distance road connecting the smallest towns; the road may or may not have a state or county route number. Scenic park roads, unimproved or unpaved roads, and industrial roads are included in this category. Most roads in the Nation are classified as A4 roads.
A5	Vehicular Trail A road in this category is usable only by four-wheel drive vehicles, is usually a one-lane dirt trail, and is found almost exclusively in very rural areas. Sometimes the road is called a fire road or logging road and may include an abandoned railroad grade where the tracks have been removed. Minor, unpaved roads usable by ordinary cars and trucks belong in category A4, not A5.
A6	Road with Special Characteristics This category includes roads, portions of a road, intersections of a road, or the ends of a road that are parts of the vehicular highway system and have separately identifiable characteristics.
A7	Road as Other Thoroughfare A road in this category is not part of the vehicular highway system. It is used by bicyclists or pedestrians, and is typically inaccessible to mainstream motor traffic except for private-owner and service vehicles. This category includes foot and hiking trails located on park and forest land, as well as stairs or walkways that follow a road right-of-way and have names similar to road names

Table 2.1. Roadway type A1 – A7 definitions. A larger definition table is given in Appendix A.

For data analysis, the GPS-based vehicle activity data were imported into the GIS environment as a separate layer, sharing the same coordinate system as the roadway network data. A query was then processed that matched the GPS data points with the closest roadway centerlines for the different roadway facility types. Figure 2.6 illustrates the mechanism of the query. In this figure, the GPS data points are shown in red. A "buffer zone" around each roadway centerline (shown in green in Figure 2.6) is created, extending perpendicularly on both sides by a specific amount. This essentially makes the roadway centerlines "thicker", as shown in blue in Figure 2.6. A query is then made on whether the GPS data points fall within these buffer zones. If they fall within a buffer zone of a particular roadway link, then they are assigned to that roadway link. If they fall outside any buffer zone, then no roadway link is assigned. If a GPS data point falls within more than one buffer zone, then the roadway link assignment is made based on the previous assignments in the GPS data path.

After this query had been processed, the selected data points are assigned to that roadway type and later exported as a new database. These steps are then repeated for all roadway types. For the road types A1, A2, and A3, the buffer zone was set to be 75 meters on each side of the road centerline. For the roadtypes A4, A5, A6, and A7, the buffer zone was set to be 30 meters on each side of the road centerline.

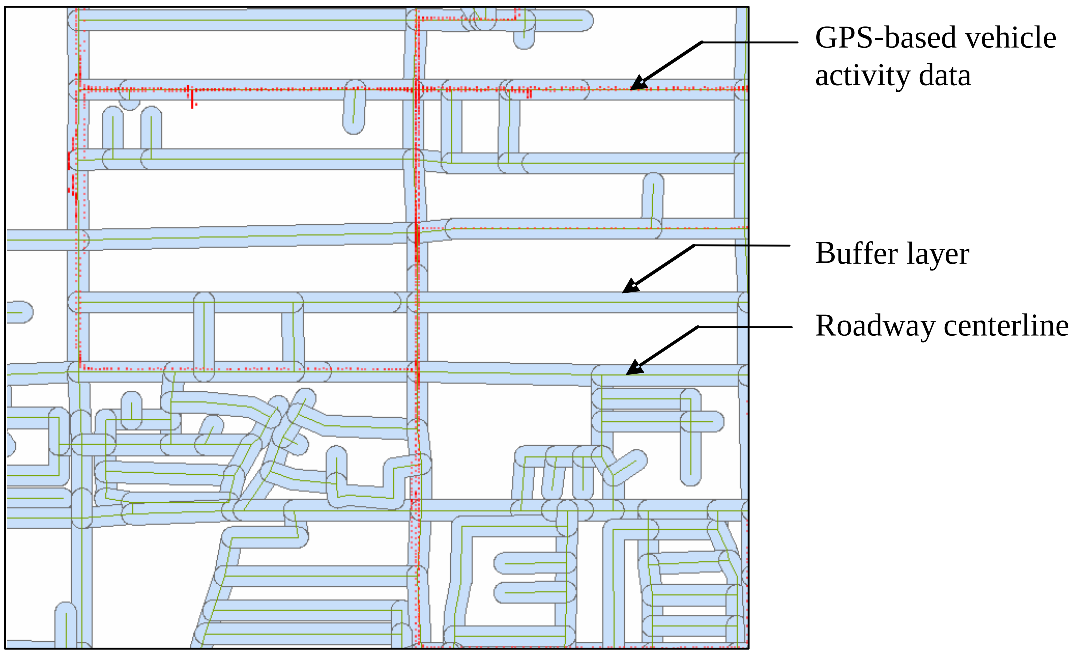

Figure 2.6. Mechanism of the spatial selection query.

It is important to note that both the CALTRANS and SCAG vehicle activity data sets that were examined were based on using standard (non-differential) GPS receivers, giving positional accuracy of approximately 20 meters. With this 20-meter accuracy, the correct match to the digital roadway network was estimated to be approximately 87% or better. Subsequent data processing was performed to eliminate the possibility of one data point being assigned to more than one roadway type (particularly where two roadway types merge). CE-CERT is currently working on other map-matching routines that have greater accuracy (see, e.g., [Du and Barth, 2006].

3. Results

For this project, the major focus was placed on two databases: the 2001 California Statewide Household Travel Survey carried out by the California Department of Transportation (referred to as the CALTRANS database); and the Southern California Association of Governments 2001 post-census travel survey (referred to as the SCAG database). These GPS-based vehicle activity databases were originally designed to provide backup information to (written) travel surveys used to determine where and why people in the study area travel. The GPS data were originally used to: 1) detect under-reporting of trips in the original telephone interview/diary retrieval process; 2) create a set of household trip rate correction factors; and 3) evaluate the accuracy of reported trip elements (e.g., trip start and finish times, locations, durations, lengths) [NuStats, 2002].

The SCAG database consisted of households randomly selected throughout the South Coast Air Basin. The CALTRANS database consisted of households selected both in Northern and Southern California.

3.1. General Characteristics and Global Statistics

The SCAG and CALTRANS databases were processed as described in Chapter 2 and put into a Microsoft ACCESS database. As the initial step in the analysis, general characteristics were extracted and global statistics were calculated. General characteristics of these databases are given in Table 3.1. The SCAG database surveyed nearly twice as many households and had significantly more vehicles that were instrumented. The CALTRANS household travel survey was done both in Northern and Southern California, whereas the SCAG database focused just on representative vehicle trips that took place in the South Coast Air Basin (SCAB). The CALTRANS survey was carried out over a longer period of time compared to the SCAG study. For both studies, household vehicles were instrumented, consisting almost entirely of light-duty passenger cars and trucks.

Statistics	CALTRANS	SCAG
Households	272	467
Total Cars	414	626
Cars/Household	1.52	1.34
Total days	283	163
Total Trips	2382	6583
Total miles of Trips	16235	28000

Table 3.1. General characteristics of the CALTRANS and SCAG databases.

Global statistics of the databases are given in Table 3.2. It can be seen that the average trip distance was slightly less in Southern California compared to the statewide average trip distance. Similarly, the average trip speed for the SCAG database was slightly lower than the CALTRANS database. In terms of trips per day, the CALTRANS database had an average trip per day per vehicle of 4.78 (trip starts) while the SCAG database had 5.2 (this is disaggregated by day of week in the next section).

It is important to note that for both studies, these GPS dataloggers were installed in the target vehicles for just a single "typical" travel day. In the SCAG database, the GPS dataloggers were installed the day prior, measurements were made during the travel day, and then removed the following day. For the CALTRANS data study, similar procedures were followed for 260 of the 272 households*. As a result, information such as "how many vehicles don't move in a day" and "first and last day installation dates" could not be extracted since day-to-day differences were not recorded for each vehicle. In order to determine that type of information, then the dataloggers must be installed for at least a week at a time, which was not done in these studies.

Statistics	CALTRANS	SCAG
Average distance per trip (miles)	6.8	4.25
Min distance per trip (miles)	0.1	0
Max distance per trip (miles)	134	146.14
Mean trips per day per household	8.55	6.82
Min. trips per day per household	1	1
Max trips per day per household	36	52
Mean trips per day per vehicle	4.78	5.205
Min trips per day per vehicle	1	1
Max trips per day per vehicle	27	37
Average speed per trip (mph)	31.48	25.3
Average time per trip (min)	8.12	12.1

Table 3.2. Global statistics of the CALTRANS and SCAG databases.

3.2. Weekday/Weekend Trip Start Differences

The number of trip starts was disaggregated by day of week to see if there are significant weekday/weekend differences. The results of these calculations are shown in Table 3.3 and Figure 3.1. Similar to other weekday/weekend studies, the number of trip starts is statistically lower on weekends compared to weekdays. However, the number of samples

* the remaining 12 households had some dataloggers installed for more than one day, however their day-to-day differences were not analyzed due to their statistically insignificant sample size.

recorded on the weekend for both the SCAG and CALTRANS datasets was significantly lower compared to the weekday.

Day	No. of trips		No. of vehicle days		Avg. trips per day per vehicle	
	CALTRANS	SCAG	CALTRANS	SCAG	CALTRANS	SCAG
Mon	441	840	71	167	6.21	5.02
Tue	612	1068	91	200	6.725	5.34
Wed	616	1531	95	263	6.484	5.82
Thurs	515	1310	94	226	5.479	5.79
Fri	382	1241	71	205	5.38	6.05
Sat	8	367	5	84	1.6	4.36
Sun	3	301	1	74	3.0	4.06

Table 3.3. Global statistics of the CALTRANS and SCAG databases.

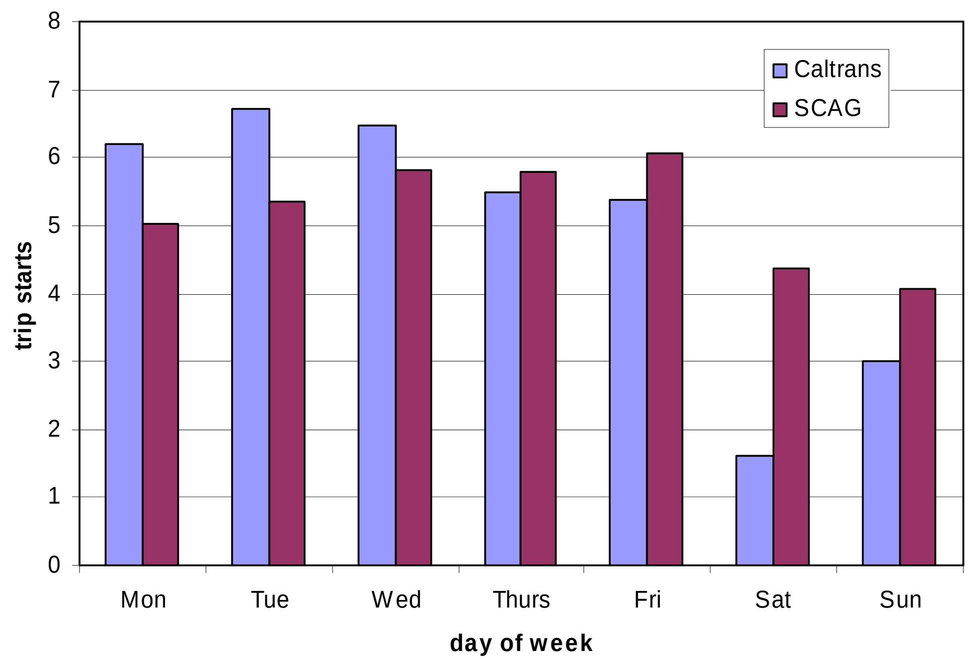

Figure 3.1. Number of trip starts by day of week, Caltrans and SCAG databases.

3.3. Trip Distance Analysis

As described in the general statistics, the SCAG database had an average trip distance of 4.25 miles and the CALTRANS database had an average trip distance of 6.8 miles. To investigate this further, a trip distance histogram was created, as shown in Figure 3.2. It can be seen that approximately 85% of all trips for both datasets were less than 10 miles. Less than 5 percent of all trips were greater than 20 miles.

Since so many trips occurred in the 0 – 10 mile bin, a histogram of these trip distances was also created, illustrated in Figure 3.3. It can be seen that according to the GPS data, most trips were very short in nature, less than one mile in distance. It is unclear whether these shorter trips are actual trips or "trips" based on how the GPS data were processed.

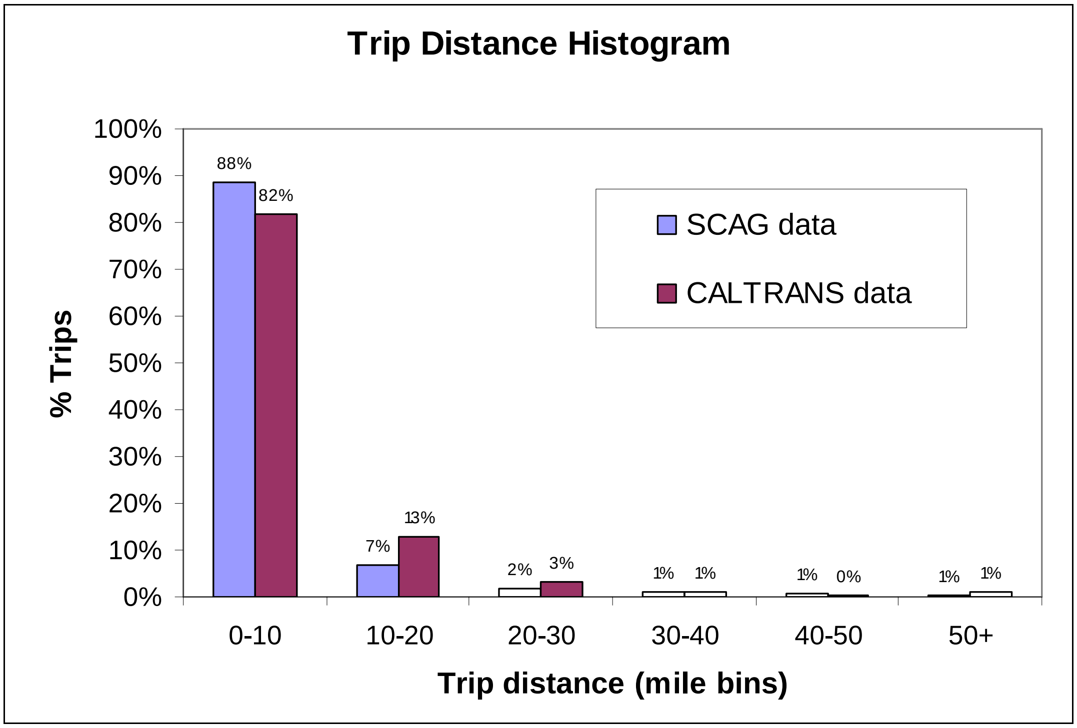

Figure 3.2. Trip Distance Histogram for the SCAG and CALTRANS databases.

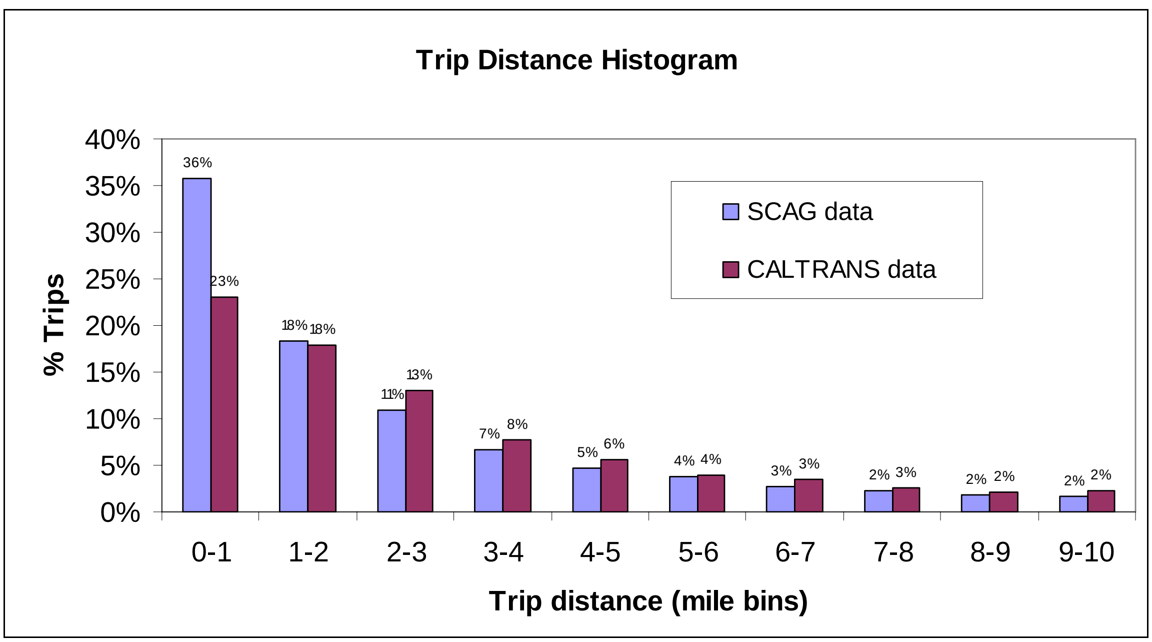

Figure 3.3. Trip Distance Histogram for the SCAG and CALTRANS databases, for distances less than 10 miles.

3.4. Trip Duration Analysis

Also stated in the general statistics, the SCAG database had an average trip duration of 12.1 minutes. Similar to Section 3.3, a trip duration histogram was created for both datasets, shown in Figure 3.4a. Approximately 80% of all trips were less than 10 minutes. Only 3.5% of all trips were greater than 30 minutes.

Again, since so many trips occurred in the 0 – 10 minute bin, a histogram of these trip durations was also created, illustrated in Figure 3.4b. It can be seen that according to the GPS data, the trip durations were spread out pretty evenly, with a peak around 3 minutes. Again, it is unclear whether these shorter trips are actual trips or "trips" based on how the GPS data were processed.

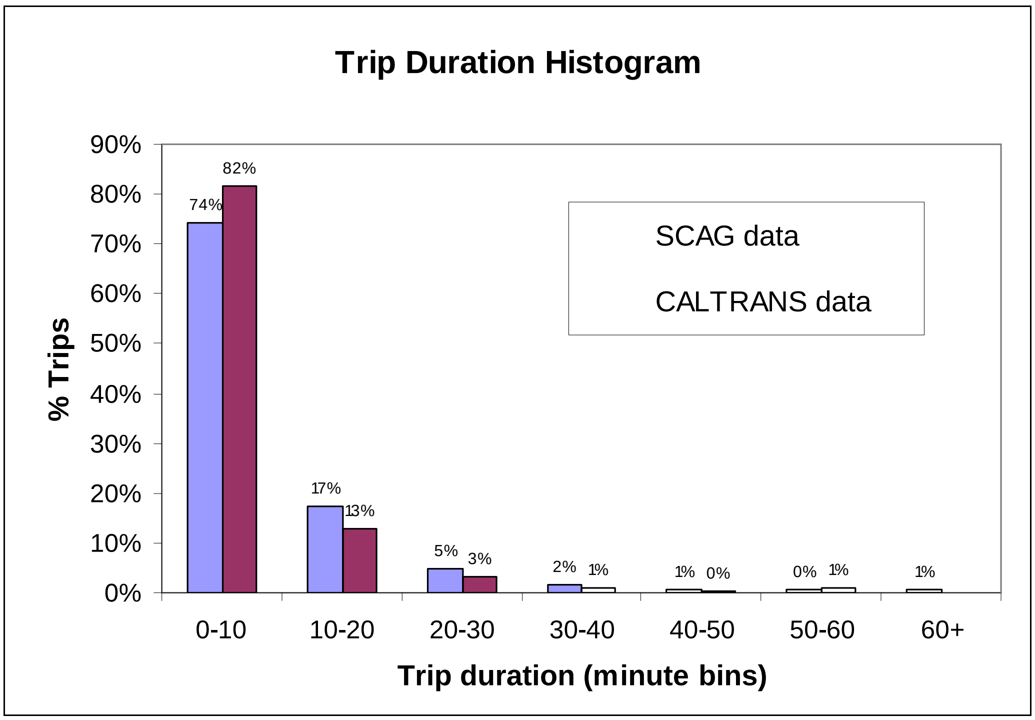

Figure 3.4a. Trip Duration Histogram of the SCAG and CALTRANS databases.

3.5. Diurnal Pattern of Trip Starts and Trip Ends

To better understand when trips begin and end, the entire SCAG database was processed, grouping trips into hourly blocks throughout the day. A histogram was created (shown in Figure 3.5) showing the diurnal patterns. Both trip starts and ends tended to follow each other rather closely. This is to be expected since the majority of the trips were quite short (10 minutes or less). In many commuting-based diurnal patterns, there is usually a distinctive A.M. peak when people commute to work and a P.M. peak when they return home. However, with other mixed-purpose trips, the peak of the trips occurs during the early afternoon (from 1PM – 4PM).

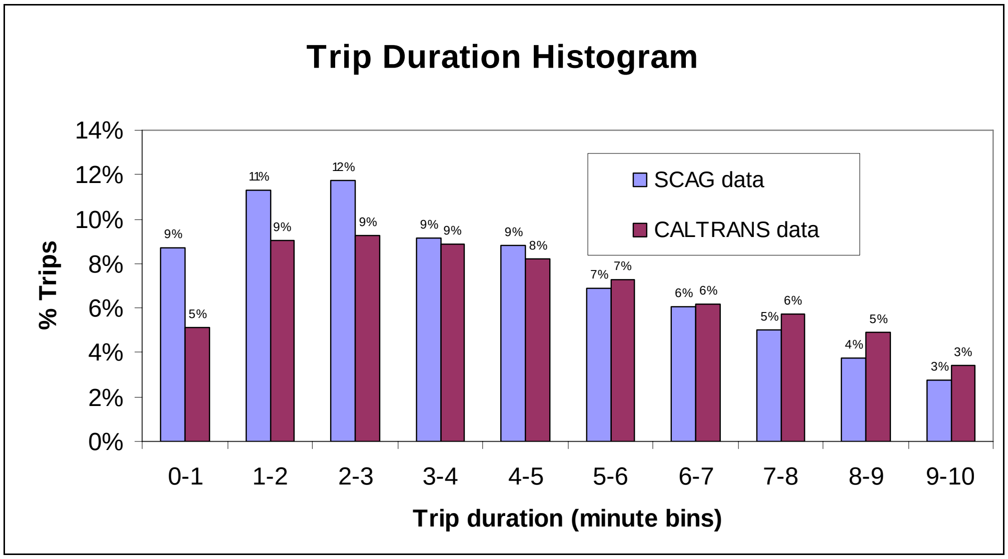

Figure 3.4b. Trip Duration Histogram of the SCAG and CALTRANS databases.

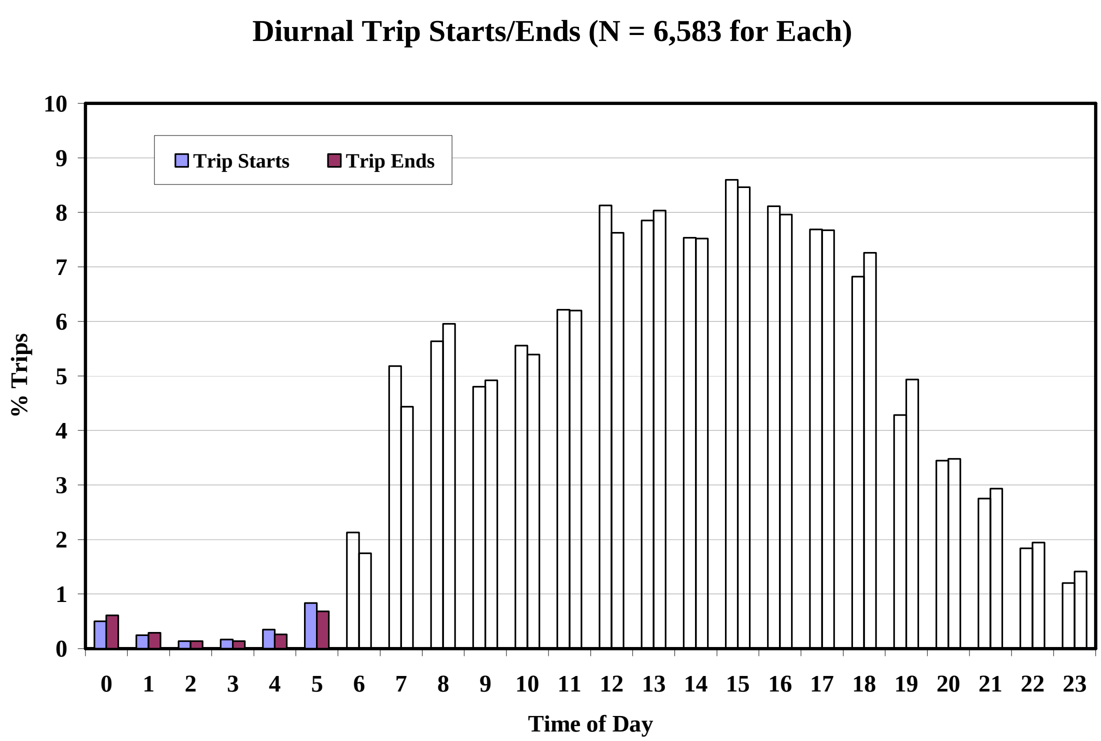

Figure 3.5. Diurnal Trip Starts/Ends of the SCAG database.

3.6. Soak Time Analysis

One of the more important parameters that was extracted from the vehicle activity databases is the soak time duration. Soak time duration is the amount of time between when a vehicle is turned off and when it is subsequently turned back on. This plays a significant role in emissions inventories, since cold engines produce much greater emissions in the first few minutes after they are first turned on after being cold. These “cold-start” events are often the majority of the emissions for the entire trip since emissions for the remainder of the trip are quite low due to effective exhaust control systems. The duration on how long an engine has been turned off plays a direct role on cold- or warm-start emissions.

A soak time duration histogram was created from the SCAG database and is illustrated in Figure 3.6. There are two peaks in this distribution. One peak occurs in the bin of 10 minutes of less (30% of the histogram). The other less pronounced peak occurs in the bin of 120 – 360 minute soak (13.5%).

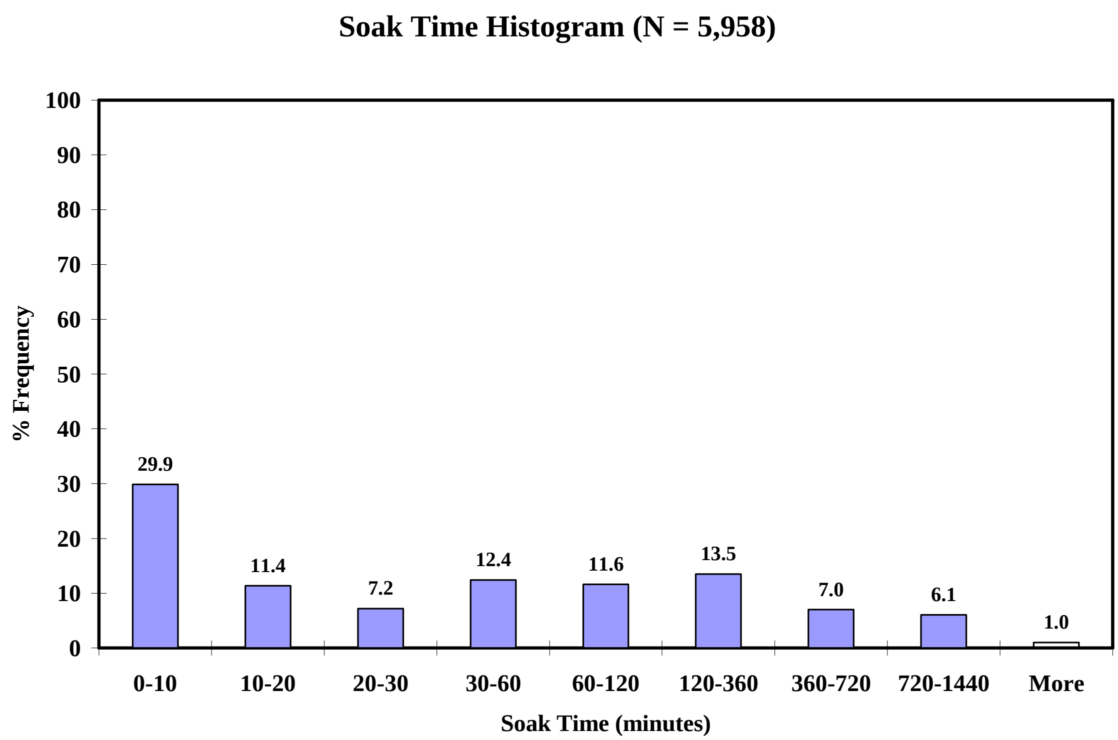

Figure 3.6. Histogram of soak times for the SCAG database.

3.7. Spatial Analysis of Trips

To help see the spatial distribution of the trips, the trip start locations were plotted geographically, as shown in Figure 3.7 and 3.8. Again, the households were recruited randomly throughout the South Coast Air Basin (SCAB). Figure 3.8 provides a zoomed-in version of the trip start locations.

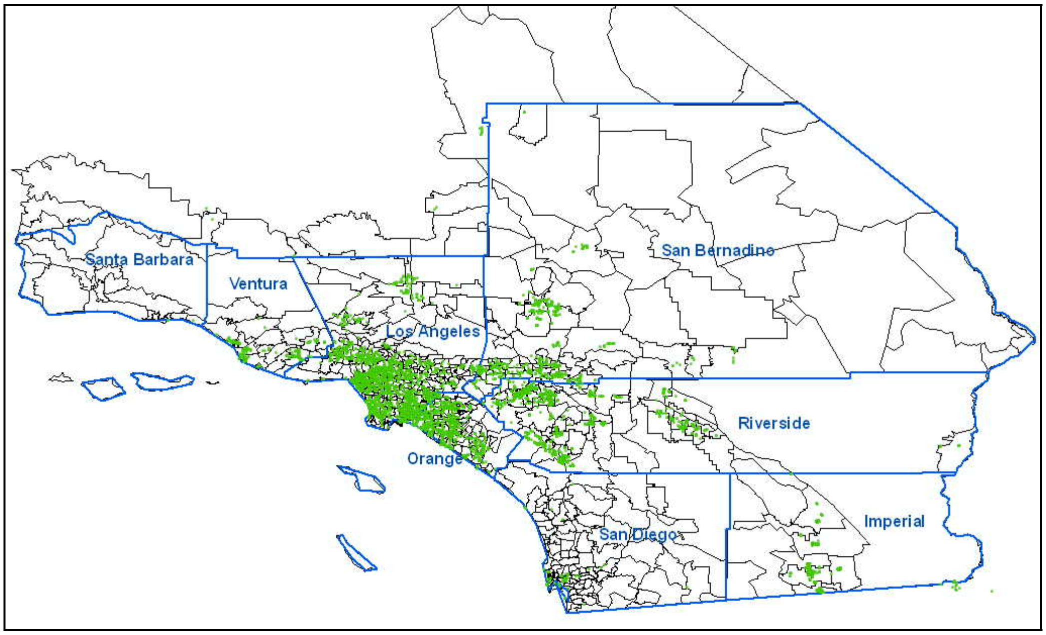

Figure 3.7. Trip start locations within the South Coast Air Basin.

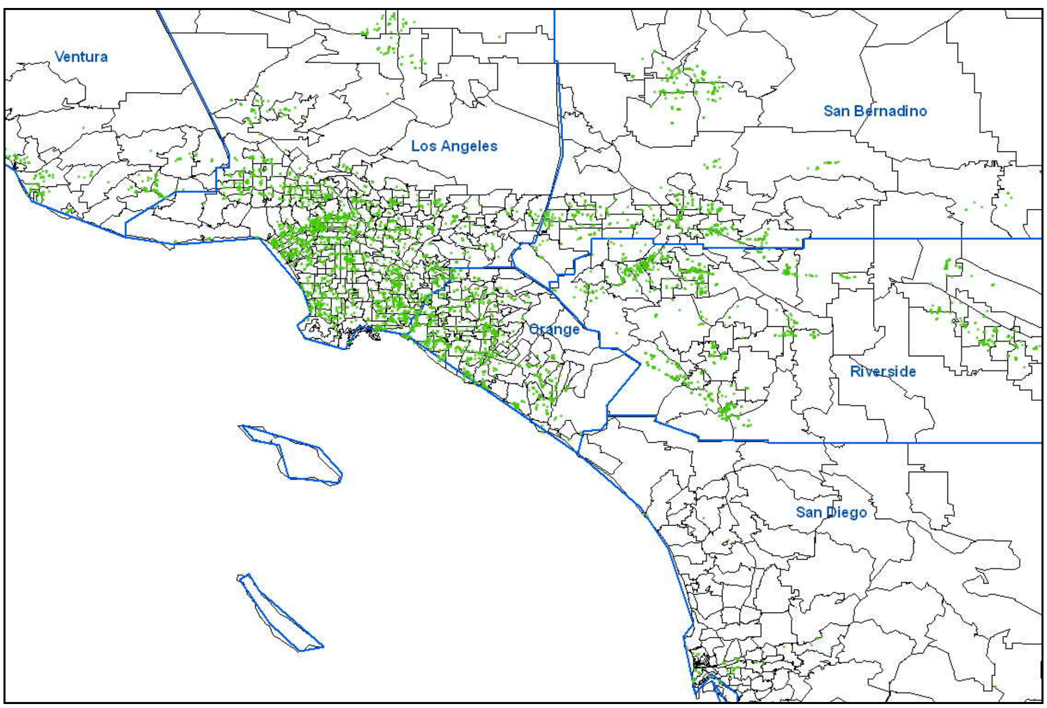

Figure 3.8. Zoomed-in version of trip start locations within the South Coast Air Basin.

Similarly, the trip end locations were plotted geographically, as shown in Figure 3.9 and 3.10 (zoomed-in version of Figure 3.9). As would be expected, the trip start and end locations tend to be quite similar when analyzed throughout a 24-hour period. Similarly for Northern California (subset of the overall CALTRANS dataset), Figure 3.11 shows the trip start locations in Northern California and Figure 3.12 shows the trip end locations in Northern California.

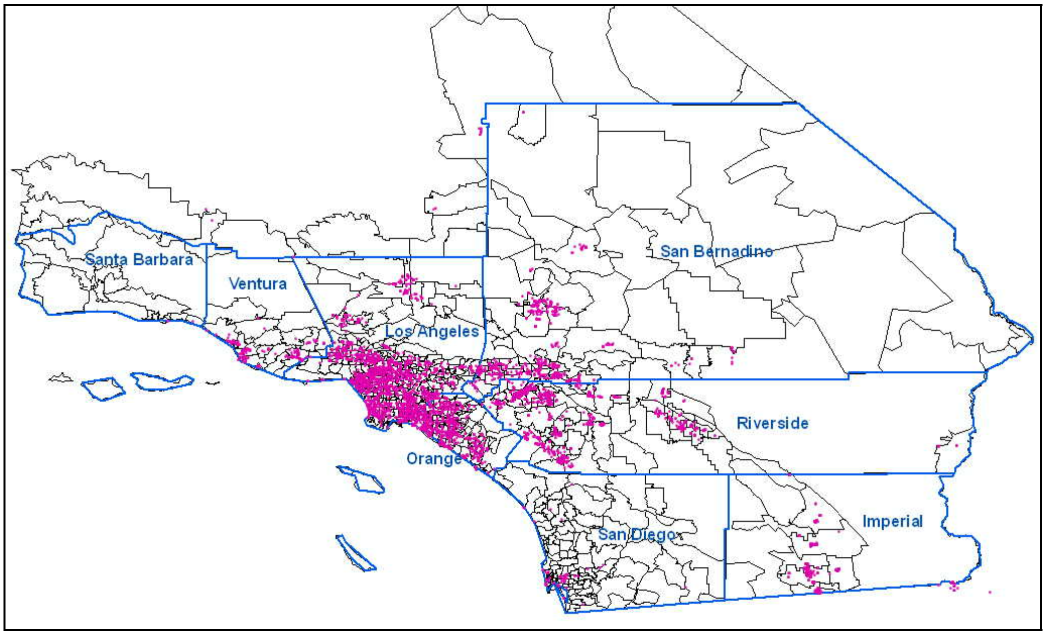

Figure 3.9. Trip end locations within the South Coast Air Basin.

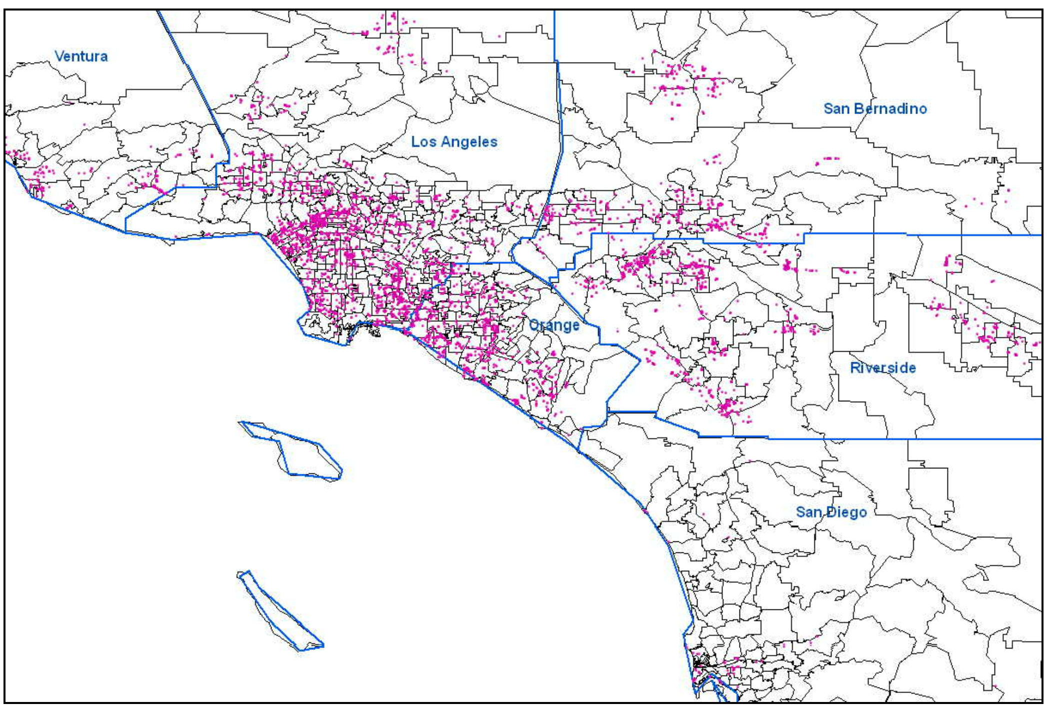

Figure 3.10. Zoomed-in version of trip end locations within the South Coast Air Basin.

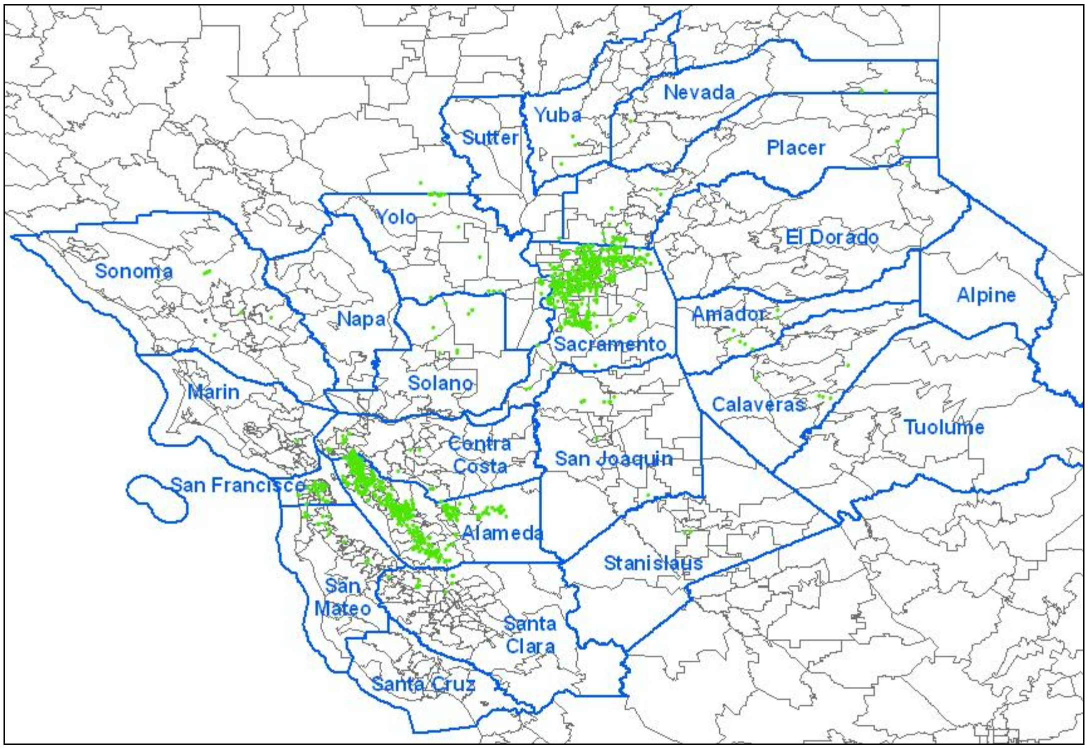

Figure 3.11. Trip start locations within Northern California.

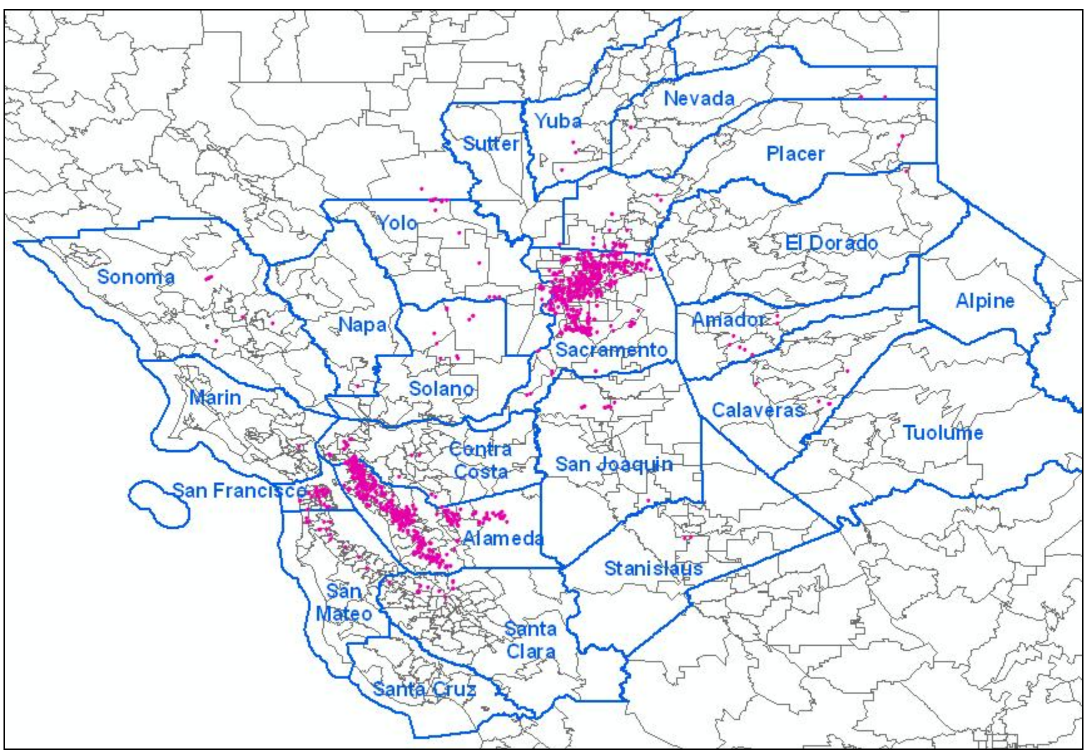

Figure 3.12. Trip end locations within Northern California.

4. Results: Roadway Facility Type Disaggregation Analysis

In order to better understand vehicle activity patterns, we have disaggregated the SCAG and CALTRANS vehicle activity databases by roadtype. As illustrated in Figure 2.4, a map-matching procedure was used to assign a roadway facility type to each data point in the vehicle trajectories. This was completed over the seven different roadway categories defined for the TIGER/Line® roadway network, listed in Table 2.1. After the vehicle trajectory database was enhanced with roadway type information, further analysis was performed as described in the following sections.

4.1. Percentage Travel by Roadway Facility Type

One of the key items to be investigated was determining the percentage of travel by roadway type. A comparison was made between travel in both Northern and Southern California. In terms of total distance, Table 4.1 and Figure 4.1 show that the majority of VMT (vehicle miles traveled) occurs on A1 roadways (primary highways with limited access), accounting for nearly 50% of all travel. In terms of differences between Northern and Southern California, there was slightly more highway travel in Southern California, based on the representative vehicle trajectories recorded in the household vehicle activity studies.

In a general sense, roadway facility types A1, A2, and A6 can be assigned as “highway” travel and roadway facility types A3, A4, A5, and A7 can be assigned as “surface street” travel. With this categorization, approximately 55% - 65% of VMT occurs on highways, and the remaining 35% -45% occurs on surface streets.

		Facility Type						
	Total Distance (mi)	A1	A2	A3	A4	A5	A6	A7
So. Cal.	6,169	2,979	158	513	1,689	2	819	9
No. Cal.	9,301	4,078	54	776	3,468	1	900	25
	Total Distance (%)	A1	A2	A3	A4	A5	A6	A7
So. Cal.	100 %	48.29%	2.55%	8.31%	27.37%	0.03%	13.28%	0.13%
No. Cal.	100 %	43.84%	0.57%	8.33%	37.28%	0.00%	9.67%	0.26%

Table 4.1. Percent of vehicle travel (miles) by roadway facility type, for Southern and Northern California.

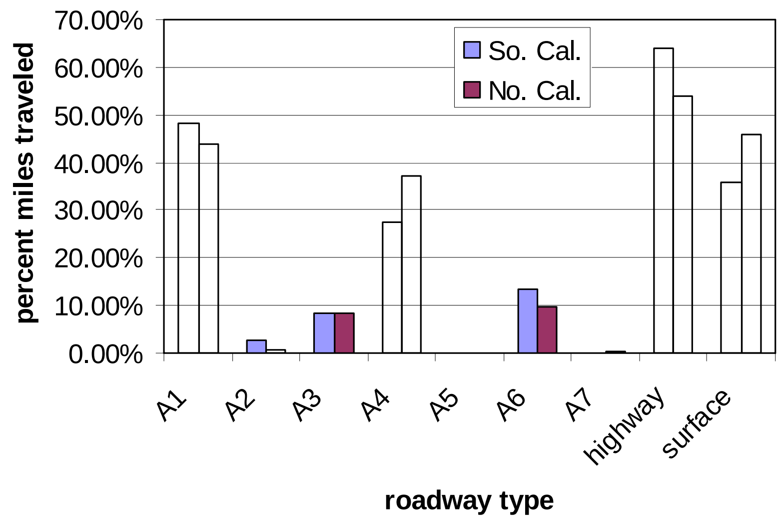

Figure 4.1. Percent of vehicle travel (miles) by roadway facility type, for Southern and Northern California.

In terms of time spent on the different roadway types, the results are illustrated in Table 4.2 and Figure 4.2. It can be seen that in this case, the majority of time spent is on the A4 roadway type (primary surface street). Southern California travel consists of slightly more time spent on the highways compared to Northern California. Overall, time spent on highways is approximately 35% - 45% on highways and 55% - 65% on surface streets.

		Facility Type						
	Total Time (s)	A1	A2	A3	A4	A5	A6	A7
So. Cal.	597,344	197,028	15,323	54,693	266,242	185	62,178	1,695
No. Cal.	918,388	273,922	3,548	73,911	494,513	81	70,046	2,367
	Total Time (%)	A1	A2	A3	A4	A5	A6	A7
So. Cal.	100%	32.98%	2.56%	9.15%	44.57%	0.03%	10.40%	0.28%
No. Cal.	100%	29.82%	0.38%	8.04%	53.84%	0.008%	7.62%	0.25%

Table 4.2. Percent of vehicle travel time by roadway facility type, for Southern and Northern California.

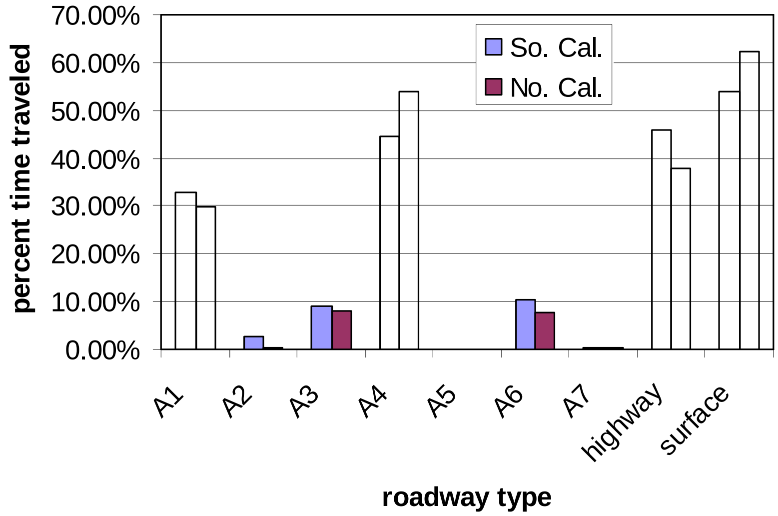

Figure 4.2. Percent of vehicle travel time by roadway facility type, for Southern and Northern California.

4.2. Operational Parameters by Roadway Facility Type

Another area of interest was to investigate the operational characteristics of the vehicle trajectories based on roadway facility type. One of the primary operational parameters is average vehicle speed. Figure 4.3 shows the average vehicle speed by roadway type. It can be seen that as expected, highway travel was generally faster than surface street travel, with an overall average of around 45 mph (35 mph for surface streets). In terms of Northern versus Southern California differences, Northern California had slightly higher highway speeds and approximately the same speeds for the surface streets.

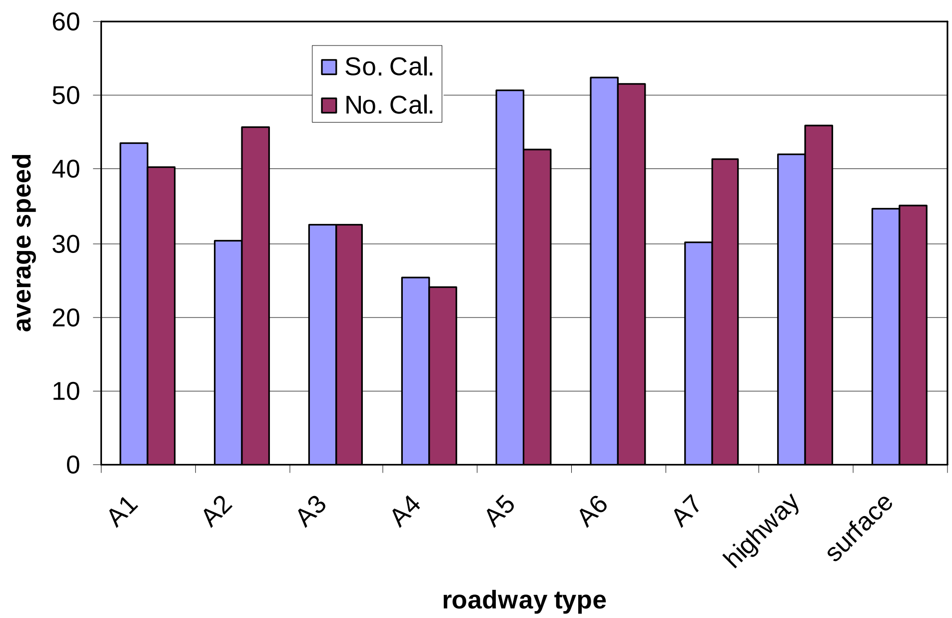

Figure 4.3. Average vehicle speed by roadway facility type, for Southern and Northern California.

In addition to average speed, we investigated other operational parameters that took into account the amount of fluctuation in the velocity patterns. Rather than just investigating average speed, we wanted to see effects of "stop-and-go" traffic, where acceleration and deceleration play a larger role. There are several parameters that can be used as a measure for this type of "noise" in the signal. They are as follows:

- ***Positive Kinetic Energy (PKE)***: this is a measure of acceleration kinetic energy per unit distance and is defined as:

  $$PKE = \frac{\sum pos(v_f^2 - v_i^2)}{x_{FI}}$$

  where the function *pos* returns only the positive values of its result, v_f and v_i are the initial and final velocity of an instantaneous interval, and x_{FI} and is the total distance traveled during the trip interval.

- ***Total Absolute second-to-second Difference (TAD):*** this measure has units of speed per mile and is defined as:

  $$TAD = \frac{\sum pos\,|\,v_f - v_i\,|}{x_{FI}}$$

  with v_f, v_i and x_{FI} having the same definitions as stated above for PKE.

- ***Coefficient of Variation (CV):*** this is defined as the standard deviation of speed (sigma) divided by the mean speed over a given interval:

  $$CV = \frac{\sigma_v}{\bar{v}}$$

- ***Acceleration Noise (ACN):*** this is defined as the standard deviation of the acceleration of the trip:

  $$ACN = \sigma_a$$

These different measures were calculated for the different roadway facility types for both Northern and Southern California. The results are given in Table 4.3. It can be seen that these parameters are generally lower for Northern California compared to Southern California. As expected, the amount of stop-and-go traffic is more pronounced on the surface streets (A3, A4, A5, and A7) compared to the highways (A1, A2, and A6).

Southern Calif.	A1		A2		A3		A4		A5		A6		A7	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Avg. Speed	43.5443	21.2768	30.2779	18.4797	32.4616	21.3729	25.4171	17.5752	50.6147	14.9300	52.5171	20.0617	30.0894	16.0085
PKE	6.0333e +003	5.4080e +003	9.6597e +003	6.6512e +003	8.5787e +003	6.4805e +003	7.9533e +003	5.2858e +003	3.2542e +003	3.5647e +003	4.9753e +003	5.5149e +003	6.3577e +003	6.5673e+003
TAD	144.0935	214.9383	299.4787	286.8277	261.4718	252.3552	261.0587	218.1525	46.7348	69.7270	95.9976	181.1698	198.1120	254.9914
CV	0.1865	0.1779	0.3222	0.1929	0.2775	0.1969	0.3151	0.1836	0.0317	0.0530	0.0881	0.1523	0.1373	0.1841
ACN	1.0435	0.9136	1.4980	0.9235	1.3545	1.0206	1.5002	1.0090	0.6419	0.8840	0.7144	0.8558	0.7329	1.0409

Northern Calif.	A1		A2		A3		A4		A5		A6		A7	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Avg. Speed	40.2750	20.5666	45.6715	22.6879	32.5022	20.7439	24.0390	15.8287	42.7605	13.6857	51.5760	20.0351	41.3844	15.4612
PKE	5.7154e +003	4.8663e +003	5.8913e +003	5.7185e +003	7.4944e +003	5.8715e +003	8.0419e +003	5.0444e +003	3.5539e +003	2.9164e +003	4.7313e +003	5.1214e +003	3.7160e +003	4.8037e+003
TAD	139.6129	198.3702	144.4084	211.0006	226.5397	234.6042	265.3283	209.1046	47.9762	43.8909	89.9621	160.6397	81.2886	184.3724
CV	0.1856	0.1788	0.1672	0.2042	0.2492	0.1941	0.3257	0.1762	0.0244	0.0243	0.0846	0.1461	0.0455	0.0997
ACN	1.0521	0.9783	1.0957	1.1340	1.2672	0.9878	1.6075	1.0112	0.3593	0.3811	0.7098	0.8664	0.4242	0.7773

Table 4.3. Acceleration noise parameters by roadway facility type, for Southern and Northern California.

4.3. Speed-Acceleration Frequency Distributions

Another way of visualizing the amount of velocity and acceleration variation within a vehicle trajectory is to create a speed-acceleration frequency distribution (SAFD). This was carried out again by roadway facility type, divided between Northern and Southern California. The results are shown in Figure 4.4. In these figures, the different contours of the histograms are shown in color.

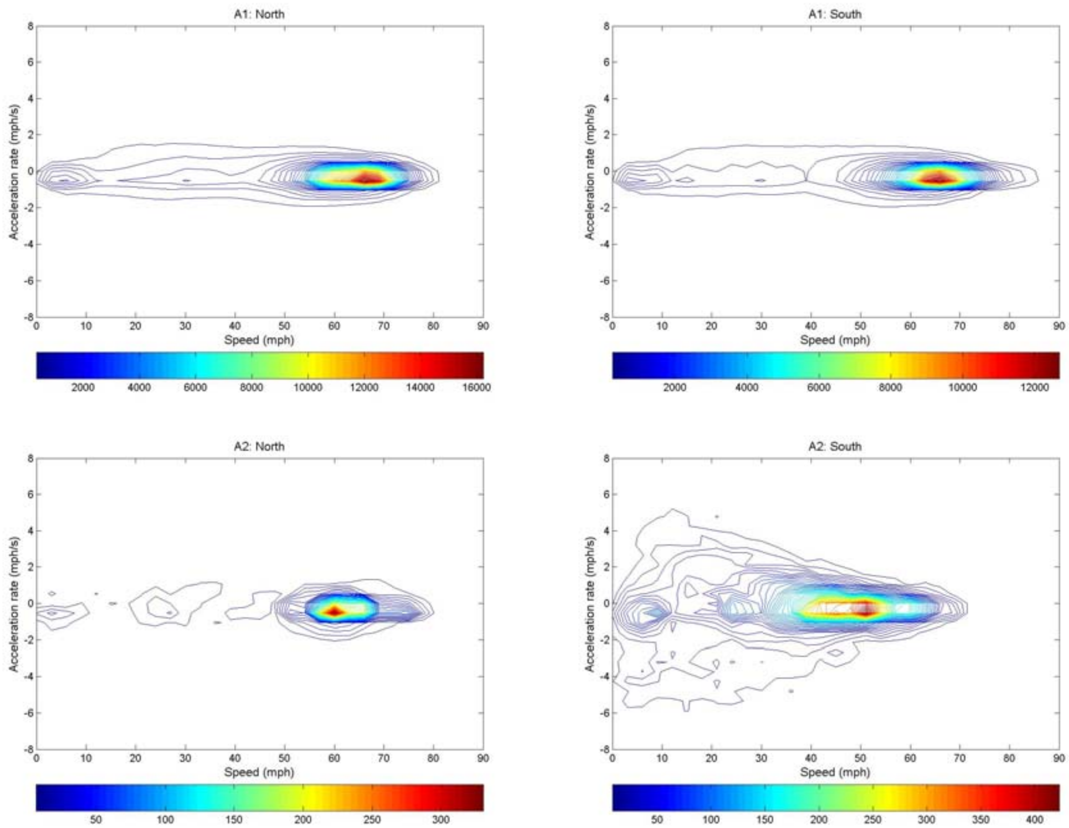

Figure 4.4a,b,c,d: Speed-Acceleration Frequency Distribution by roadway facility type, for Southern and Northern California.

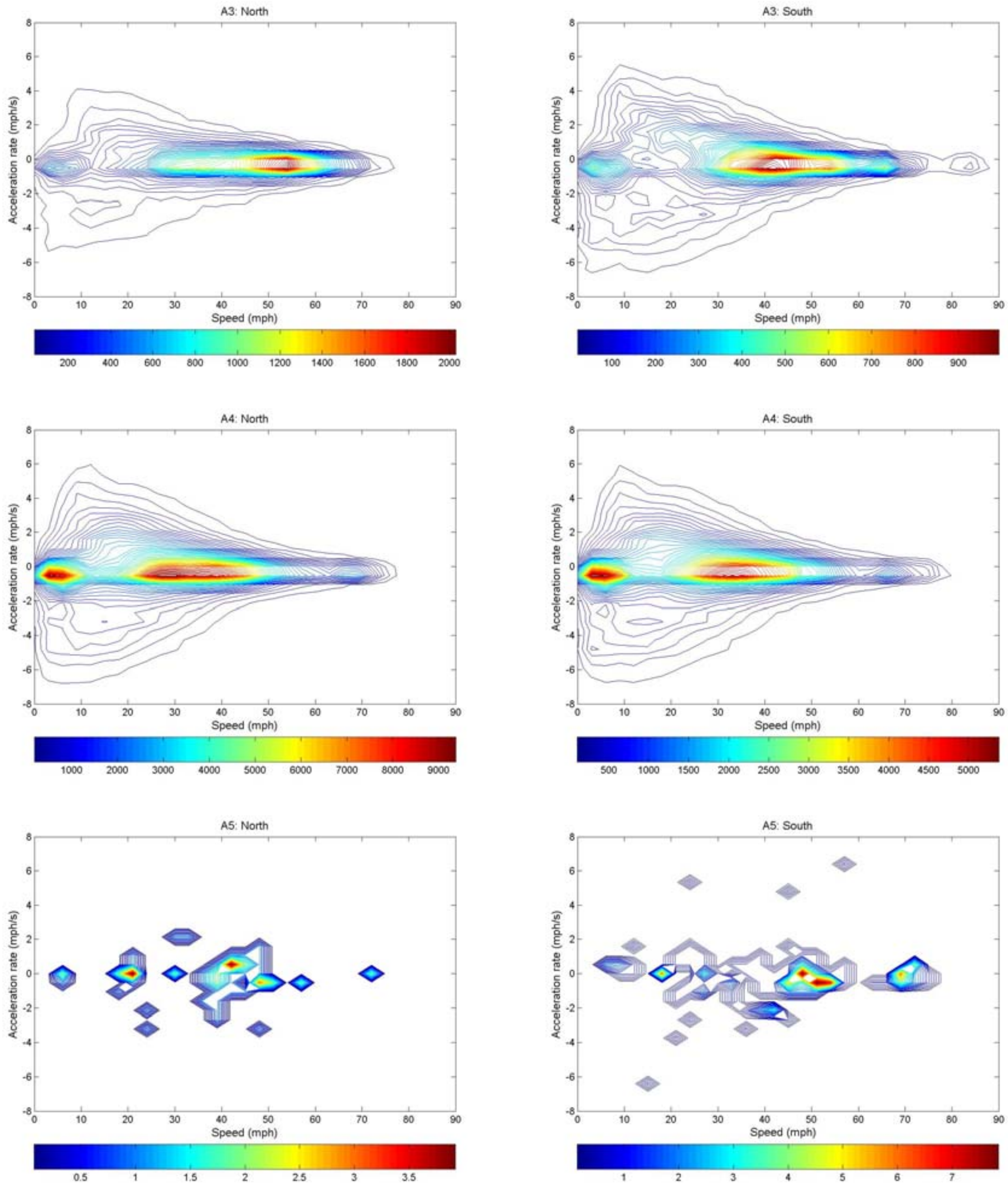

Figure 4.4e,f,g,h,i,j: Speed-Acceleration Frequency Distribution by roadway facility type, for Southern and Northern California.

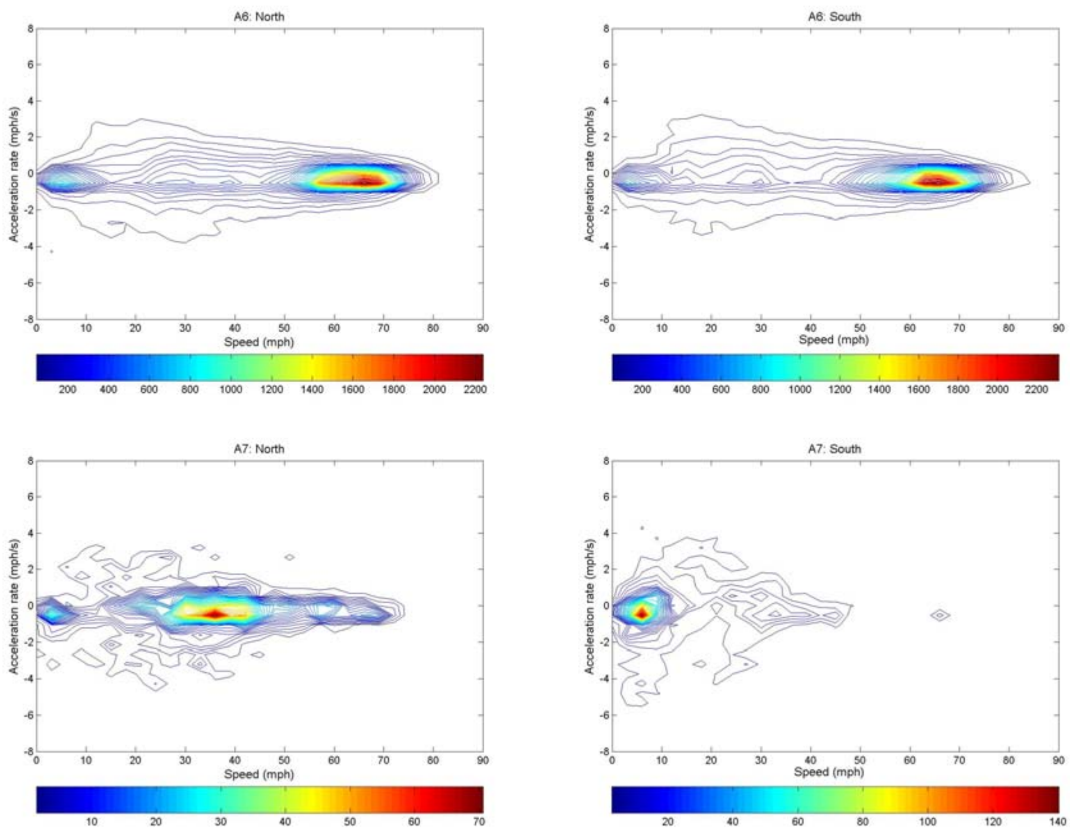

Figure 4.4k,l,m,n: Speed-Acceleration Frequency Distribution by roadway facility type, for Southern and Northern California.

It can be seen that for the highways, most of the travel occurs in the higher speed bins with little variation in the acceleration. The surface roadways on the other hand show significant acceleration variation at lower speeds, as is expected. In terms of differences between Southern and Northern California, the A2 roadway type showed slightly more variation in Southern California compared to Northern California.

4.4. Wavelet Parameter Analysis

Another powerful time-series analysis tool is the use of wavelet transforms. Wavelet transforms were developed primarily in the 1980's as a tool to divide data or functions into different frequency components. Each time-series component can then be examined at different scales. Wavelet transforms are well known for their good localization of both time and frequency instead of a frequency-only technique like a discrete Fourier transforms (DFT). There has been a significant amount of activity in recent years applying wavelet techniques to transportation data (see, e.g., [Jiang, 2004; Ping, 2005; Yu, 2005]).

Although Fourier transform is well known for its frequency analysis of a signal, it has no way of "localizing" frequencies for different parts of the time series. A wavelet transform can help us solve this problem and it is well known for its good localization of both time and frequency.

Basically, there are two kinds of wavelet transform. The first one is continuous wavelet transform (CWT). Let $x(t) \in L^2(R)$, where $L^2(R)$ is the set of all functions f such that the integral of f^2 over the whole real line is finite, and $\phi(t)$ is the mother wavelet function, then the CWT of *x(t)* is

$$WT_x(a,\tau) = \frac{1}{\sqrt{a}} \int x(t)\phi^*(\frac{t-\tau}{a})dt =< x(t), \phi_{a\tau}(t) > \qquad (4.1)$$

where a is the scale factor and τ is the translation factor.

CWT is calculated by computing the correlation between a signal and a continuously shifted and continuously scaled version of the mother function. Therefore, much redundancy exists in a CWT. At the same time, for most functions, the CWT has no analytical solutions and therefore need to be numerically calculated by a computer. Moreover, in many practical applications, the signal is sampled time sequences. Therefore, a discrete wavelet transform (DWT) is more preferable. An important issue in any transform scheme is the question of reconstruction from the transform domain. It turns out that it is possible to reconstruct a signal from its wavelet decomposition. DWT is quite suitable for discrete signal processing, for example, in speech, image, and time sequence processing. There are several ways to present the DWT: from a filter-bank theory approach or from a multiresolution analysis (MRA) point of view. We briefly provide the MRA approach (see [Hrenandez & Weiss, 1996] and [Huhtala et al., 1999] for further details).

MRA describe mathematically the process of studying signals at different scales. MRA represents the whole space $L^2(R)$ by a sequence of embedded subspaces for an intelligent choice of appropriate subspaces for an application to get a compromise between accuracy and computation complexity. MRA studies the property of a sequence of subspace V^j and W^j, $j \in Z$, which approximate $L^2(R)$ by satisfying:

$$\cdots, V^1 = V^0 \oplus W^0, V^2 = V^1 \oplus W^1, \cdots, V^{j+1} = V^j \oplus W^j \cdots$$

where:

- $Union(V^j)_{j\in Z} = L^2(R)$ (union of all V^j is the whole space);
- $\cap_{i\in Z} V^j = O$ (intersection of all V^j is empty);
- ($V^j \perp W^j$ V^j is orthogonal to W^j); and
- $W^j \perp W^{j'}$, $(j \neq j'$ W^j is orthogonal to $W^{j'}$).

The Haar scaling function is defined as:

$$\phi(t) = \begin{cases} 1 & t \in [0,1) \\ 0 & t \notin [0,1) \end{cases} \qquad (4.2)$$

and $\phi^j(t) = 2^{j/2}\phi(2^j t - i)$, $j = 0,1,...and \quad i = 0,1,...,2^j - 1$.

The vector space V^j is defined as $V^j = span\{\phi_i^j\}_{i=0,1,\ldots,2^j-1}$. It is straightforward to see that $V^j \subset V^{j+1}$. The Haar wavelet function is defined as

$$\varphi(t) = \begin{cases} 1 & t \in [0,1/2) \\ -1 & t \in [1/2,1) \\ 0 & otherwise \end{cases} \tag{4.3}$$

and $\varphi_i^j(t) = 2^{j/2}\varphi(2^j t - i) \quad j = 0,1,...and \quad i = 0,1,...,2^j - 1$.
The vector space W^j is defined as $W^j = span\{\varphi_i^j\}_{i=0,1,...2^j-1}$ such that $W^j \subset V^{j+1}$.

By straightforward mathematical computation, we find that the Haar basis has this very important property:

$$V^{j+1} = V^j \oplus W^j$$

In order to compute a Haar wavelet transform, we let $f \in R^N$, where $N = 2^n$, and expand this signal into $V^0 \oplus W^0 \oplus W^1 \oplus \cdots \oplus W^{n-1}$ spanned by the basis vectors. This is accomplished by first expanding f into $V^{n-1} \oplus W^{n-1}$. The first step can be described as the following matrix-vector multiplication $f_1 = W_1 f$ where

$$W_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 & & & & & & \\ & & 1 & 1 & & & & \\ & & & & \ddots & \ddots & & \\ & & & & & & 1 & 1 \\ 1 & -1 & & & & & & \\ & & 1 & -1 & & & & \\ & & & & \ddots & \ddots & & \\ & & & & & & 1 & -1 \end{bmatrix}_{2^n \times 2^n}$$

It can be seen that the first half rows correspond to the basis $\phi_0^{n-1}, \phi_1^{n-1}, \cdots, \phi_{2^i-1}^{n-1}$ which span V^{n-1} and the last four rows correspond to the basis vector $\varphi_0^{n-1}, \varphi_1^{n-1}, \cdots, \varphi_{2^j-1}^{n-1}$ which spans W^{n-1}.

Step two is to expand f into $V^{n-2} \oplus W^{n-2} \oplus W^{n-1}$. The second step can be described by the matrix-vector multiplication as $f_2 = W_2 f_1$.

$$W_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 & & & & & & & & & \\ & & \ddots & \ddots & & & & & & & \\ & & & & 1 & 1 & & & & & \\ 1 & -1 & & & & & & & & & \\ & & \ddots & \ddots & & & & & & & \\ & & & & 1 & -1 & & & & & \\ & & & & & & 1 & & & & \\ & & & & & & & 1 & & & \\ & & & & & & & & \ddots & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \\ & & & & & & & & & & & 1 \end{bmatrix}_{2^n \times 2^n}$$

This can be generalized to:

$$W_n = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 & & & \\ 1 & -1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}_{2^n \times 2^n}$$

and $f_n = W_n W_{n-1} \cdots W_1 f$, where f_n is the wavelet coefficient.

One method to deal with the problem of high dimensionality is to reduce the dimensionality by combining features. Linear combinations are preferable because they are simple to compute. Principal components analysis (PCA) is one kind of such linear combination methods and it seeks a projection that best represents the data in a least square sense. PCA finds components that are useful for representing data. The main application of PCA is to reduce the dimensionality of data while keep as much information as possible. Since an eigenvalue decomposition of the covariance matrix is necessary, the computation complexity can be high.

In the case of using vehicle velocity trajectories as the time-series data, there are two key properties: (1) a vertical shifting variance; and (2) a trend variance. We assume that average speed of the short vehicle velocity trajectory is an important indicator of the road type and congestion level. Therefore, roadtype and congestion classification is highly sensitive to vertical shifting. Tilting the sequence also affects roadtype classification greatly. We find that the Haar wavelet transform is good for roadtype and congestion classification since its first coefficient is the average of the data. Haar coefficients are also sensitive to the general trend because tilting the sequence affects all the Haar coefficients. In this project, we first use discrete wavelet transform for feature extraction. The resulting feature vectors have many useful properties. They give us a good representation of the natural features of the original time series, not only at small scale, such as sharps peaks, but also at large scale, such as wide mountains. After the wavelet transform, PCA can be applied to wavelet subbands to extract features further and therefore reduce the dimensionality.

In this project, we use the Haar DWT transform as the feature extractor. Subsequently, PCA was applied to further reduce the dimensionality.

It is well known that the magnitudes of wavelet coefficients of a signal at each resolution level are proportional to the corresponding energy in the signal. In Figure 4.5, the energy concentration is plotted as a function of resolution. Fifty sequences for each class were used in this analysis. Each sequence has a dimension of 128 seconds. Resolution "0" corresponds to the low frequency component, and resolutions 1 through 7 correspond to the high frequency components. The low frequency component and resolution 1 through 7 together make the Haar wavelet transform [Chan, 1999].

In the CALTRANS database, the A2, A5, A6 and A7 data are limited in size, therefore there isn't enough data for these groups to have a meaningful wavelet analysis. Therefore, we consider wavelet energy concentration analysis for the three groups: A1, A3, and A4 for both Southern and Northern California. Again, the Southern California vs. Northern California data comparison is based on the CALTRANS database.

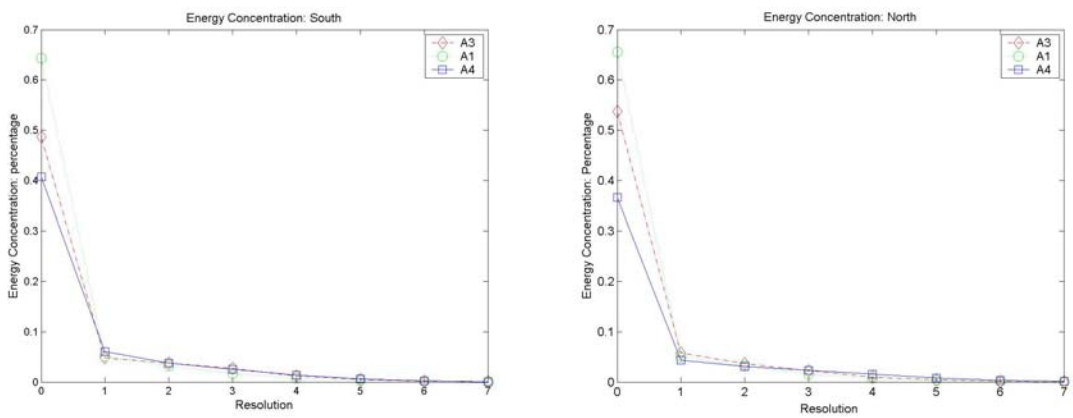

Figure 4.5. The energy concentration per coefficient vs. resolution, both in the Southern and Northern California.

For this analysis, concentration is defined as follows:

$$concentration = \frac{Sum_{amp\_in\_a\_resol}}{Sum_{all} \times number_{coef\_in\_a\_resol}}$$

where $Sum_{amp\_in\_a\_resol}$ is the sum of all the values of amplitudes with a resolution; Sum_{all} is the sum of all the amplitudes of all resolutions; and $number_{coef\_in\_a\_resol}$ is the number of coefficients within a resolution.

It can be seen from Figure 4.5 that for these three roadway types, they all have their energy concentrated in the first few scales. This verifies our original hypothesis. It is interesting to note that within the three types of roads, A1 concentrates its energy in the first few coefficients much more than A3 and A4. There is a slight difference between the A3 and A4 regarding energy concentration with A3 being higher.

4.5. Total Daily Trip Analysis

Another area of investigation was disaggregating the vehicle activity by the total miles driven by a particular vehicle in a single day. This is different from just analyzing single trip distances; instead it accounts for all trips made in a day, summing up the total mileage. This type of analysis can be important for determining the range requirements for electric, hybrid-electric, and plug-in hybrid electric vehicles.

For this analysis, the SCAG database was disaggregated into groups of driving distance, ranging from 0 – 10 miles, 10 – 20 miles, etc. all the way out to 100+ miles traveled (in a single day). These results were further disaggregated by roadway facility type. The total results of this analysis are given in Appendix B. As a summary, the percent of total miles driven by these mileage groups are shown in Figure 4.6. It can be seen that the groups that travel the furthest primarily conduct their travel on the freeways. In contrast, the shorter daily travel groups tend to use the surface streets. In general, the majority of all daily travel is less than 30 miles per day, and this occurs primarily on the surface streets.

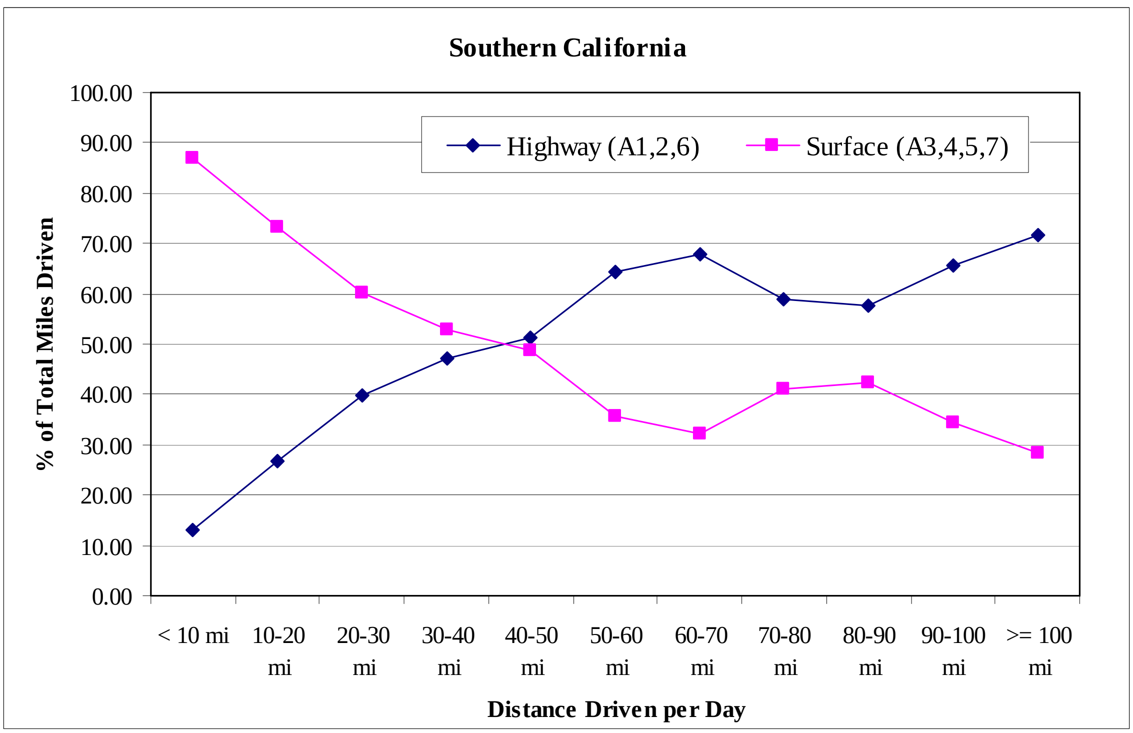

Figure 4.6: Distribution of total miles driven per day for the SCAG database.

5. Conclusions and Recommendations

The overall goal of this research project was to acquire two recently completed GPS-based vehicle activity dataset carried out in California. One of these datasets came from the Southern California Association of Governments post-census travel survey (carried out in 2001) and the other from the California Department of Transportation (CALTRANS) statewide household travel survey program, also carried out in 2001. These datasets were original created to provide backup information to written travel surveys. Since these data acquisition programs were carried out across a wide range of representative households, it was felt that under further analysis, these data could provide useful information in characterizing vehicle activity for creating accurate mobile source emissions inventories.

The data were acquired, processed, and put into a useable database format. Subsequent data analysis focused on vehicle start patterns, general trip characteristics, and detailed operational activity disaggregated by roadway facility type. From this analysis, several general conclusions can be made:

- The average distance per trip was relatively short (4 to 6 miles), with trips slightly shorter in Southern California compared to Northern California.

- The number of trips per day per vehicle was approximately 5 for both data sets.

- The average trip duration for the datasets was around 8 to 12 minutes, slightly longer in Southern California.

- The household-based datasets showed that there were little differences of travel from Monday – Friday, however on Saturdays and Sundays, the trips were significantly reduced.

- An analysis of the diurnal trip patterns for the two household datasets did not show a typical commute pattern with a distinctive AM morning peak and a PM afternoon peak. Instead, most activity peaked during the early afternoon in a single mode distribution.

- An analysis of the soak time periods of the vehicles showed a two-mode distribution, where one peak occurring for 10 minutes of less (30% of the distribution and the other less pronounced peak occurring in the range of 120 – 360 minutes (13.5% of distribution).

- After disaggregating the dataset by roadway facility type, it was seen that approximately 55% - 65% of VMT occurs on freeways, and the remaining 35% - 45% occurs on surface streets.

- In contrast, trip time spent on highways is approximately 35% - 45% while for surface streets, it was approximately 55% - 65%.

- Average speeds were significantly higher on highways (as expected) compared to surface streets. Northern California had slightly higher speeds overall.

- A number of speed-acceleration parameters and speed-acceleration frequency distributions were evaluated across the vehicle activity databases; as expected, surface streets displayed greater speed-acceleration fluctuation compared to highway travel.

Recommendations:

In terms of next steps, some of the results from this study can be used to update the vehicle activity portion of the EMFAC emissions inventory process. Based on this preliminary analysis, several additional tasks could follow:

- It is now possible to create roadway facility specific emission factors for the different kinds of driving that occur on the different road types. This can be accomplished by taking the driving snippets from the corresponding facilities and running them through a modal emissions model (weighted for a specific fleet). This would then allow for a link-based emissions inventory process where activity is measured on a link-by-link basis then multiplied by the corresponding emissions factor.

- Similarly, it is possible to create representative “driving cycles” that correspond to specific roadway facility types. These driving cycles could also be used to create facility-specific emission factors through a real-world test program.

- Now that appropriate analysis tools have been developed for processing GPS-based vehicle activity datasets, it is now possible to carryout additional vehicle activity studies at a fairly low cost. It is not very expensive to put in GPS dataloggers into representative vehicles and use those vehicles as “probes” to determine traffic and activity conditions. Of particular interest would be a truck travel pattern study.

- Hybrid electric and plug-in hybrid electric vehicle energy management strategies can be optimized using these vehicle activity data. Since the vehicle activity is representative of real-world driving patterns, the energy management of a charge-sustaining strategy or zero emission range can be optimized based on the data sets.

References

Barth, M., F. An, T. Younglove, C. Levine, G. Scora, M. Ross, and T. Wenzel. (1999) “The Development of a Comprehensive Modal Emissions Model”. Final report submitted to the National Cooperative Highway Research Program, November, 1999, 255 p.

Barth, M., C. Malcolm, T. Younglove, N. Davis, and J. Lents. (2002). Developing Link Emission Factors In The Southern California Air Basin”. *Proceedings of the 12th CRC On-Road Vehicle Emissions Workshop,* San Diego, CA, April, 2002.

Bhat, C., S. Srinivasan, and S. Bricka (2005) “Conversion of Volunteer-Collected GPS Diary Data into Travel Time Performance Measures: Literature Review, Data Requirements, and Data Acquisition Efforts”, Federal Highway Administration/Texas Transportation Research Report, Number FHWA/TX-05/0-5176-1.

Chan, K. P. and A. Fu (1999) “Efficient Time Series Matching by Wavelets,” *Proc. Int’l Conf. Data Eng.*, 1999.

Davis, N., J. Lents, M. Osses, N. Nikkila, and M. Barth, (2005) “Development and Application of an International Vehicle Emissions Model”, to be presented at the 2005 Transportation Research Board’s Annual Meeting, Washington D.C., January 2005.

Du, J. and M. Barth (2006) “Bayesian Probabilistic Vehicle Lane Matching for Link-Level In-Vehicle Navigation”, IEEE Intelligent Vehicles Conference, Tokyo Japan, June, 2006.

ESRI (2006) “TIGER/Line® Database”, accessed July 2006, see http://arcdata.esri.com/data/tiger2000/tiger_statelayer.cfm?sfips=06.

Farrell, J. and M. Barth (1999) *The Global Positioning System and Inertial Navigation: Theory and Practice*, ISBN-0-07-022045-X, McGraw-Hill Publishers.

Gammariello, R., and J.R. Long, (1996) “Development of Unified Correction Cycles” - CRC Sixth Annual On-Road Vehicle Emissions Workshop, San Diego, CA, 1996.

Guensler, R., S. Washington, and W. Bachman. (1998) “Overview of the MEASURE Modeling Framework”. Pp. 51-70 in *Transportation Planning and Air Quality III: Emerging Strategies and Working Solution*. American Society of Civil Engineers, ISBN 0-7844-0355-4, 1998.

Hart, C., J. Koupal, and R. Giannelli (2002) “EPA's Onboard Analysis Shootout: Overview and Results”, EPA Technical Report # 420-R-02-026, October 2002.

Hernandez, E. and G. Weiss, (1996) *A first Course on Wavelets*. CRC Press, 1996.

Huhtala, Y., J. Karkkainen, and H. Toivonen (1999) “Mining for Similarities in Aligned Time Series Using Wavelets, *SPIE Conference on Data Mining and Knowledge Discovery,* SPIE vol. 3695, pp. 150-160.

Jiang, X. and H. Adeli (2004) "Wavelet Packet-Autocorrelation Function Method for Traffic Flow Pattern Analysis," *Computer-Aided Civil and Infrastructure Engineering*, Vol. 19, 2004, pp. 324-337 .

Koupal, J. et al., (2002) "Draft Design and Implementation Plan for EPA's Multi-Scale Motor Vehicle and Equipment Emission System (MOVES)", U.S. EPA Technical Report #420-P-02-006, October 2002.

Malcolm, C., T. Younglove, M. Barth and N. Davis. (2002) "Mobile Source Emissions: Analysis of Spatial Variability in Vehicle Activity Patterns and Vehicle Fleet Distributions". *Transportation Research Record No. 1842*, pp 91 – 98, Journal of the Transportation Research Board, National Academy of Science.

NuStats (2002) "Year 2000 Post-Census Regional Travel Survey for Southern California Association of Governments", final report submitted to the Southern California Association of Governments, August 2002.

Ping, Y. (2005) "Linear prediction of traffic volume", *Proceedings of 2005 Transportation Research Board Annual Meeting*, Washington D.C., January 2005 also at http://www.missouri.edu/~sunc/TRB/presentations.html.

Sonoma Technology, Inc., (2004), "Collection and Analysis of Weekend/Weekday Emissions Activity Data in the South Coast Air Basin", Final Report submitted to the California Air Resources Board, May, 2004.

TRB NCHRP Synthesis 301 (2001) "Collecting, Processing, and Integrating GPS data into GIS: A Synthesis of Highway Practice", NCHRP Synthesis 301, Transportation Research Board National Research Council.

U.S. EPA, (1993) "Federal Test Procedure Review Project: Technical Report", EPA Technical Report # 420-R-93-007, May 1993.

Yu, L. (2005) "Wavelet-Based Determination of Aggregation Levels for ITS Data", *Proceedings of 2005 Transportation Research Board Annual Meeting*, Washington D.C., January 2005 also at http://www.missouri.edu/~sunc/TRB/presentations.html.

Appendix A: TIGER/Line File Census Feature Class Codes (CFCC)

Road Class	Description
A1	Primary Highway With Limited Access Interstate highways and some toll highways are in this category (A1) and are distinguished by the presence of interchanges.
A11	Primary road with limited access or interstate highway, unseparated
A12	Primary road with limited access or interstate highway, unseparated, in tunnel
A13	Primary road with limited access or interstate highway, unseparated,underpassing
A14	Primary road with limited access or interstate highway, unseparated, with rail line in center
A15	Primary road with limited access or interstate highway, separated
A16	Primary road with limited access or interstate highway, separated, in tunnel
A17	Primary road with limited access or interstate highway, separated, underpassing
A18	Primary road with limited access or interstate highway, separated, with rail line in center
A2	Primary Road Without Limited Access Includes nationally and regionally important highways that do not have limited access as required by category A1. It consists mainly of US highways, but may include some state highways and county highways that connect cities and larger towns. A road in this category must be hard-surface (concrete or asphalt). It has intersections with other roads, may be divided or undivided, and have multi-lane or single-lane characteristics
A21	Primary road without limited access, US highways, unseparated
A22	Primary road without limited access, US highways, unseparated, in tunnel
A23	Primary road without limited access, US highways, unseparated, underpassing
A24	Primary road without limited access, US highways, unseparated, with rail line in center
A25	Primary road without limited access, US highways, separated
A26	Primary road without limited access, US highways, separated, in tunnel
A27	Primary road without limited access, US highways, separated, underpassing
A28	Primary road without limited access, US highways, separated, with rail line in center
A3	Secondary and Connecting Road Includes mostly state highways, but may include some county highways that connect smaller towns, subdivisions, and neighborhoods. The roads in this category generally are smaller than roads in Category A2, must be hard- surface (concrete or asphalt), and are usually undivided with single-lane characteristics. These roads usually have a local name along with a route number and intersect with many other roads and driveways
A31	Secondary and connecting road, state highways, unseparated
A32	Secondary and connecting road, state highways, unseparated, in tunnel
A33	Secondary and connecting road, state highways, unseparated, underpassing
A34	Secondary and connecting road, state highways, unseparated, with rail line in center
A35	Secondary and connecting road, state highways, separated
A36	Secondary and connecting road, state highways, separated, in tunnel
A37	Secondary and connecting road, state and county highways, separated, underpassing
A38	Secondary and connecting road, state and county highway, separated, with rail line in center
A4	Local, Neighborhood, and Rural Road A road in this category is used for local traffic and usually has a single lane of traffic in each direction. In an urban area, this is a neighborhood road and street that is not a thoroughfare belonging in categories A2 or A3. In a rural area, this is a short-distance road connecting the smallest towns; the road may or may not have a state or county route number. Scenic park roads, unimproved or unpaved roads, and industrial roads are included in this category. Most roads in the Nation are classified as A4 roads.
A41	Local, neighborhood, and rural road, city street, unseparated
A42	Local, neighborhood, and rural road, city street, unseparated, in tunnel

A43	Local, neighborhood, and rural road, city street, unseparated, underpassing
A44	Local, neighborhood, and rural road, city street, unseparated, with rail line in center
A45	Local, neighborhood, and rural road, city street, separated
A46	Local, neighborhood, and rural road, city street, separated, in tunnel
A47	Local, neighborhood, and rural road, city street, separated, underpassing
A48	Local, neighborhood, and rural road, city street, separated, with rail line in center
A5	Vehicular Trail A road in this category is usable only by four-wheel drive vehicles, is usually a one-lane dirt trail, and is found almost exclusively in very rural areas. Sometimes the road is called a fire road or logging road and may include an abandoned railroad grade where the tracks have been removed. Minor, unpaved roads usable by ordinary cars and trucks belong in category A4, not A5.
A51	Vehicular trail, road passable only by 4WD vehicle, unseparated
A52	Vehicular trail, road passable only by 4WD vehicle, unseparated, in tunnel
A53	Vehicular trail, road passable only by 4WD vehicle, unseparated, underpassing
A6	Road with Special Characteristics This category includes roads, portions of a road, intersections of a road, or the ends of a road that are parts of the vehicular highway system and have separately identifiable characteristics.
A60	Special road feature, major category used when the minor category could not be determined
A61	Cul-de-sac, the closed end of a road that forms a loop or turn-around
A62	Traffic circle, the portion of a road or intersection of roads forming a roundabout
A63	Access ramp, the portion of a road that forms a cloverleaf or limited-access interchange
A64	Service drive, the road or portion of a road that provides access to businesses, facilities, and rest areas along a limited-access highway; this frontage road may intersect other roads and be named
A65	Ferry crossing, the representation of a route over water that connects roads on opposite shores; used by ships carrying automobiles or people
A7	Road as Other Thoroughfare A road in this category is not part of the vehicular highway system. It is used by bicyclists or pedestrians, and is typically inaccessible to mainstream motor traffic except for private-owner and service vehicles. This category includes foot and hiking trails located on park and forest land, as well as stairs or walkways that follow a road right-of-way and have names similar to road names
A70	Other thoroughfare, major category used when the minor category could not be determined
A71	Walkway or trail for pedestrians, usually unnamed
A72	Stairway, stepped road for pedestrians, usually unnamed
A73	Alley, road for service vehicles, usually unnamed, located at the rear of buildings and property
A74	Driveway or service road, usually privately owned and unnamed, used as access to residences, trailer parks, and apartment complexes, or as access to logging areas, oil rigs, ranches, farms, and park lands