TOUR-BASED ANALYSIS OF MULTI-DAY GPS DATA

Peter Stopher, Institute of Transport and Logistics Studies, The University of Sydney

Yun Zhang, Institute of Transport and Logistics Studies, The University of Sydney

Qingjian Jiang, Parsons Brinckerhoff, Australia

ABSTRACT

The purpose of this research is to understand travel patterns by applying tour-based analysis and using sociodemographic variables to characterise travel patterns to explore new opportunities of developing activity-based and tour-based models. The data used in this research is from an Australian panel where 47 households provided GPS data for a period of 28 days, with a total of 89 persons.

This paper presents the results of a basic tour analysis of the above data, which includes the distribution of tours and trips per day, tour duration and the starting times by trip purpose, followed by a summary of important considerations when dealing with tour-based data. We further introduce an extended tour classification, based in part on work initially done by O'Fallon and Sullivan (2009), where a set of twelve tour classifications are put forward, based on a hierarchy of trip purposes of work, education, shopping, and other. With the application of the new tour classification, we present the findings concerning the relationship between the purpose of the tours, the composition of the tours (simple or complex tours) and the characteristics of the day (weekend or week day), sociodemographic characteristics, such as employment or education status, and the stages in the family life cycle.

Overall, for one of the first times, multi-day GPS data were used for tour based analysis. The findings of this paper enrich the current understanding of tour patterns and provide significant insight into multi-day travel, which may yield fruitful explanations of tour-based travel to build improved tour choice models.

BACKGROUND

Over the past several years, the Institute of Transport and Logistics Studies (ITLS) has been collecting a considerable amount of data using Global Positioning System (GPS) devices. The devices are capable of collecting data on a second-by-second basis as people travel and are easily portable, so can be carried by people whether they are walking, riding a bicycle, riding a public transport vehicle, or riding in or driving a car. The devices used to collect most of the data are illustrated in Figures 1 and 2, with Figure 2 being the latest version of the devices.

Figure 1: First GPS Device (Neve)

Figure 2: Second GPS Device (Atmel)

There are four major data collection activities that ITLS has pursued. First, a panel of households was set up in 2005, comprising nominally 200 households. In these households, each member of the household over the age of 14 was asked to carry the first GPS device (Figure 1) with them wherever they went for 7 days. Households were selected randomly, using a parcel-based sampling method from a GIS of land use of the region. The same households were asked to repeat this in 2006 and 2007. Because of attrition, the panel make-up changed slightly during the course of the study. A somewhat larger sample was used in 2006 for reasons that are explained elsewhere (Stopher et al., 2009). Second, a panel of 50 households, also sampled in the same manner as a random sample form the parcel-based GIS, was recruited in 2006. In these households, each person over the age of 14 was asked to carry the first GPS device with them for 28 days. These households repeated this exercise six months later. Following that, they were combined with the first panel, but asked to carry the first devices with them for 15 days rather than 28 in the third wave in 2007. Third, a panel of 120 households, drawn randomly from several states in Australia, was recruited in 2007. The source listing for these samples were households that had been contacted in previous TravelSmart interventions, and included both households that engaged in TravelSmart and households that did not. This panel is using the second GPS device and has been asked to carry the devices for 15 days each year. This survey is continuing until 2012, with three years of data collected so far (2007, 2008, and 2009). Some of these 120 panel members include panel members from the other two panels. Finally, a sample of over 3,000 households is being asked to carry GPS devices as part of a GPS-only household travel survey in Ohio, USA (Giaimo et al., 2010; Stopher and Wargelin, 2010). In this survey, all members of each sampled household over the age of 12 are asked to carry the second type of device (Figure 2) with them for 3 days. A subsample of respondents is then asked to complete a prompted recall survey that provides additional data on the travel and also allows verification of the results of GPS data processing. Households are being sampled randomly from an address-based listing.

In addition to collecting the GPS data, the panels and the HTS sample in Ohio have been asked to complete sociodemographic data forms for each person and household, and also to provide data on the vehicles available for use by household members. The sociodemographic data includes address data on the workplaces of each person in the household, the educational establishments attended by members of the household, and the two most frequently used grocery stores for each household. Further, data have been assembled in a GIS of the street system for each locality where respondents live, work, and travel, the public bus routes and bus stops, and, for the first two panels, the land use of each parcel in the urban areas where respondents reside.

ITLS has also developed software over the past several years to process these GPS data (Stopher et al., 2008a). The software initially uses several heuristics to split the second-by-second traces into what are assumed to be identifiable trips. The rules include procedures that usually separate trips legs that use different modes, such as a walk to the bus stop, followed by a ride on the bus, followed by a walk to the destination. This is done by looking at the sustained speeds of movement, as well as identifying the bus portion of the travel by its coincidence with a GIS of bus routes and a beginning and ending point that coincide with a GIS of bus stop locations. Following the initial identification of trips, the procedure requires a visual check of the results to make sure that, as far as possible, the trips look sensible and to pick up any possible stops that may have been missed in the automated process. Following the visual checking procedure and any edits to the trip file, the next step is to identify the mode used on each identified trip.

The software that ITLS has developed uses the 85th percentile speed, rates of acceleration and deceleration, and location of the path relative to roads, bus routes, and rail lines on a GIS to identify the mode. The software will classify each trip segment to walk, bicycle, car, bus, or rail. At the moment, there is no procedure available to identify if the car user is a driver or passenger, but work is proceeding on a way of classifying the number of household members occupying a privately owned vehicle. Finally, the software identifies trip purpose. This is done partly from the collection of several addresses that is part of the data collection process. The addresses collected are those of the home, the workplaces of each person in the household who works, the educational establishments attended by any members of the family, and the two most frequently used grocery stores. In addition, if a GIS of parcel land use is available, this is used. The other information that is used to classify purpose is the frequency of visits over the period of days for which GPS data are collected and the duration of those visits.

While none of these software procedures are completely accurate, tests to date suggest that the accuracy level is very high. It must be kept in mind that standard self-report data do not provide complete accuracy on any of these attributes of travel, because people typically give incorrect responses on some trips and also often provide only partial or even wrong addresses for the places they visit. Information on addresses visited, purpose of trips, mode of travel, etc. are also sometimes missing from diary records. From tests that ITLS has conducted, trip identification is believed to be accurate to within about ±2 percent, mode to within about ±12 percent and purpose to within about ±20 percent (Stopher et al., 2010). However, these figures have yet to be completely verified and current work by ITLS is hoped to provide more complete validation of this.

Standard Modelling Approaches

In general, modelling of human travel behaviour has been based almost entirely on a one-day snapshot of each household's travel, gained from a self-report travel survey, such as a diary. For instance, Bowman and Ben-Akiva (2001) presented an integrated activity-based model based on a twenty-four hour household travel diary survey. Vovsha, Bradley and Bowman (2005) presented an overview of the development of activity/tour-based models from 1995 to 2005, as well as developments before 1995, almost all of which were based on one-day self-report data.

In standard approaches, data may be used at either the household or the person level. Assumptions are made that the data, which may be collected over a period as long as a year (even three years in the case of some continuous travel surveys, such as the Sydney Travel Survey (Battelino and Peachman, 2001)) can be combined and treated as though all travel days are representative of an average travel day throughout the year. Further, it is then an assumption of the modelling that the data may be pooled from all sampled households and persons to provide the estimation data for determining the parameters of some set of travel-demand, activity, or tour-based models. In most models, socio-demographic characteristics of the travellers may appear as additional variables in the model, although some models assume that the coefficients of travel-related attributes are themselves a function of the sociodemographic characteristics of the traveller and the models are segmented by these characteristics.

A consequence of this type of data for model estimation is that the models must usually be of a form where each model produces a probabilistic estimate of an aspect of travel behaviour, while being based on the observations of what decisions were actually made on a particular day by a particular individual or household. In other words, the modelling paradigm is to take binary data that indicates either that a certain behaviour out of some choice set occurs, or that it does not occur, and convert those binary data into probabilistic models. In addition, these data also provide only static information to input to the travel models – there is no information pertaining to the dynamics of effects on travel behaviour. One of the only departures from these approaches of using one-day self-report data is that of Axhausen, et al. (2002), who analyzed the rhythms in travel behaviour based on self report data collected from a six-week continuous travel diary project named Mobi*drive*.

ADVANTAGES OF THE ITLS GPS DATA

Considering the background data on how models of travel behaviour have been built in the past, the first and most obvious advantage that is offered by GPS data pertains to the fact that GPS data are typically collected for a number of days from each respondent. Moreover, such multi-day data is not subject to the fatigue effects usually encountered with more conventional data collection procedures. Typically, if diaries are used for multiple days, the level of reporting completeness and accuracy tends to drop as the period of time lengthens. This is a result of the tedium of the self-reporting survey, especially when it comes to reporting travel that may seem to the respondent to be quite repetitive of previous days that have already been reported. Indeed, one might expect two things to happen with multi-day reporting through a conventional self-report survey. First, one would expect that the respondent would tend to omit reporting more of the short trips and other travel that the respondent considers not interesting, as time goes by. Even in a two-day diary, Stopher et al. (2006) found that there was a marked fall off in reporting completeness on the second day of the diary and this has been reported by others in two-day and longer surveys (Pas, 1986; Hanson and Huff, 1988; Axhausen, et al., 2002; Axhausen et al., 2007; *inter alia*). Second, one would expect that repetitive trips, such as travel to and from work, would be reported identically from day to day, even when there were in fact variations in the travel, because copying the same data from one day to the next would reduce the amount of effort and thinking required in a multi-day diary. In other words, missing out some trips and repeating the details of other trips without reporting accurately on real variations would both be mechanisms for reducing respondent burden on a multi-day dairy.

Neither of these effects is present in a GPS multi-day survey. Each day of the survey, the respondent simply has to carry the GPS device with him or her and remember, at the end of each day, to keep it charged. There is no relationship between burden (which is very slight anyway, compared to a multi-day diary) and the amount of travel data reported in a given day. Hence, fatigue effects will not be present in multi-day GPS data. The fact, then, that multiple days of data can be collected rather readily with GPS devices then opens up a new possibility. With multiple days of data, it would be possible to construct probabilities of particular travel events occurring on any given day. For example, the GPS multi-day survey may provide data that shows that a particular person went grocery shopping on two days out of seven. This could then be converted into a daily probability of shopping of 0.286, and estimation of model parameters can be done using simpler modelling structures, because there is now an observation of a probability and also measurement of appropriate and relevant characteristics of the person, the household, and the shopping travel. Moreover, if a particular person is found to travel to work by car on three days a week, to work from home on one day and to use public transport on one day, all of which are weekdays, then one could assign mode probabilities to the days of work, as well as an overall probability of making a trip to work.

In the data collected by ITLS, most of the GPS data provides at least seven days of daily travel data, while some of the data provide 15 and even 28 days of data. Full household and person demographics were collected. Thus, along with the travel characteristics that can be deduced from the GPS records, and the demographic data, there are considerable potentials for developing models of travel behaviour from the GPS data.

Another advantage of the data that ITLS has collected is the fact that these data are mainly from panels. A panel is defined here as repeat measurement of the same individuals and households on two or more occasions. The panels actually provide data annually, with some panel members now having provided data for up to five years. This provides the possibility to examine the dynamics of travel behaviour, and especially to see how both external events and changes within the household affect travel behaviour. This is information that has rarely been available previously, because only few panels have ever been established in transport and this is the first panel with multi-day data of the magnitude of 7 or more days per person and household.

Thus, the GPS data provide very accurate information about the times, duration, and locations of travel, along with very detailed route information, and provide this for multiple days and for multiple years. To illustrate the nature of the data that are available, a few statistics are useful. These are provided in the next section of this paper.

A 28-DAY PANEL WAVE

Using one wave of data from the panel of 50 households, the following statistics provide a brief idea of the available data. In the panel, there were 47 households that provided good data, comprising 89 respondents who carried GPS devices for part or all of the 28-day period. Theoretically, that would provide 2,492 person days of data, but there were 1,515 person days of travel data including verifiable no-travel days. From those 1,515 person days of travel data, there are 2,274 tours, averaging approximately 26 tours per respondent. There are also 612 person days recorded on which no travel took place. The reason that these add to more than 2,492 person days is that a number of people had actually recorded data on more than 28 days, with some providing data for as many as 32 days.

In this research, a tour is defined as all of the travel and activities that take place from when a person leaves home until that person returns to home. In some definitions of tours, there are sub-tours defined, especially based on workplaces. However, in this work, the concept of a sub-tour is not used. Thus, in the statistics reported above and throughout the rest of this paper, a tour is all of the travel and activities from home back to home again. A person will make a second tour if he or she leaves home a second time in the day and performs another sequence of travel and activities, returning back to home again. It is also important to keep in mind in reviewing the following statistics that measurement includes weekend days as well as weekdays and that the average 28-day period of measurement will include 8 weekend days.

In the 28-day data under study, almost 28 percent of person days consisted of only one tour. The average number of tours performed per day was 1.75. The mean duration of a tour was 2 hours, 50 minutes, and 6 seconds. The median was 107.6 minutes (1 hour, 47 minutes, and 34 seconds). It was found that 38.7 percent of the tours comprised just two trips, while the mean number of trips per tour was 2.63 and the median was 2. There were also 293 one-trip tours, most of which would be activities like walking the dog, or jogging in the neighbourhood, and all were, as expected, home back to home trips. This averages to less than one trip per person per week. Figure 3 shows the distribution of the number of tours per day, while Figure 4 shows the distribution of the number of trips per tour for these data. As might be expected, Figure 3 shows that few people made more than four tours per day, with the vast majority making only one or two tours per day and the maximum being 8 tours in a day.

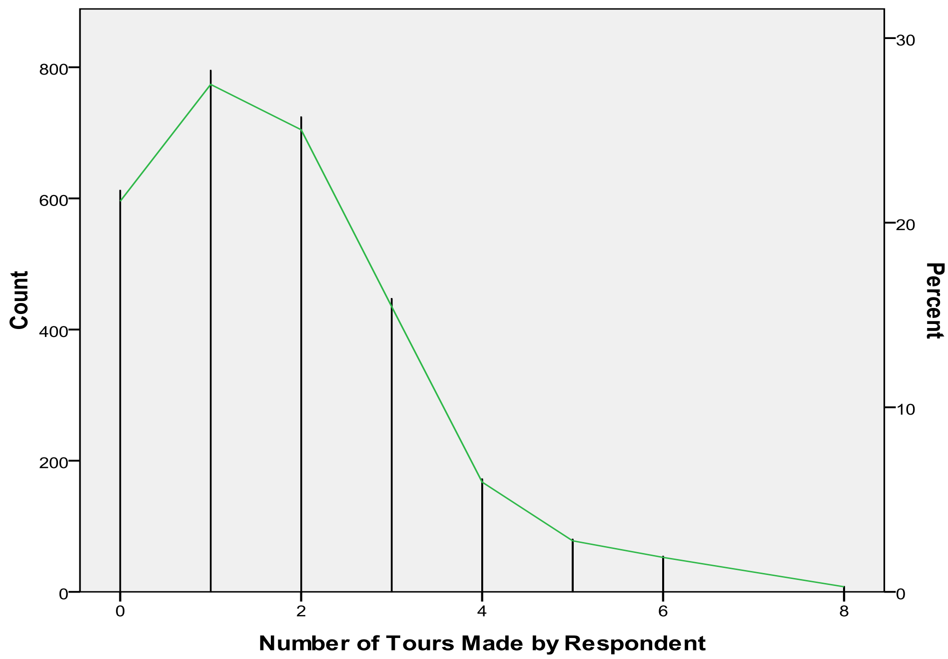

Figure 3: Distribution of Tours per Day

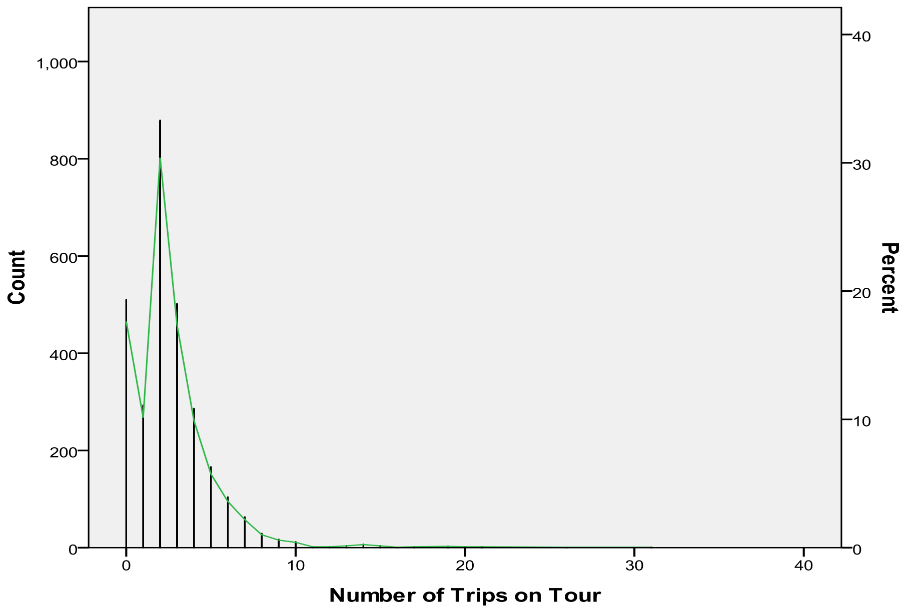

Figure 4: Distribution of Trips per Tour

Figure 4 shows the dominance of 2 and 3 trip tours, but also shows that a small fraction of tours involved as many as 8, 9, and 10 trips, with the maximum number of trips on a tour being 31. The distribution of tour travel time durations is shown in Figure 5. This distribution shows a peak in travel time durations of about 20 to 30 minutes, but with a very long tail that extends to well beyond 8 hours in duration. Figure 6 shows the distribution of dwell times within tours, i.e., the amount of time spent in activities away from home during a tour. There is a preponderance of zero values for the 28 percent of tours that were one-trip tours and therefore had no dwell time in the tour. However, the dwell times show a decreasing number with increasing values and there are few dwell times above category 60 which represents 10 hours.

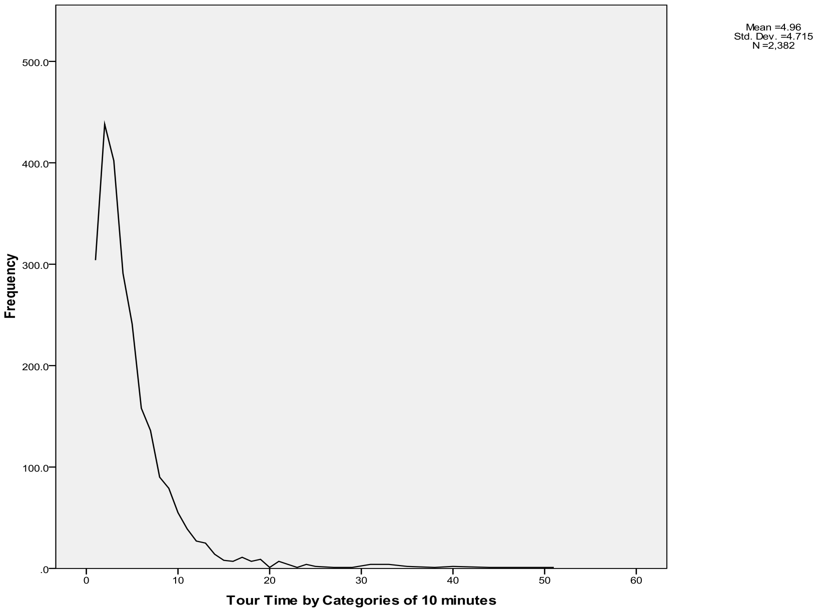

Figure 5: Distribution of Tour Durations in Minutes

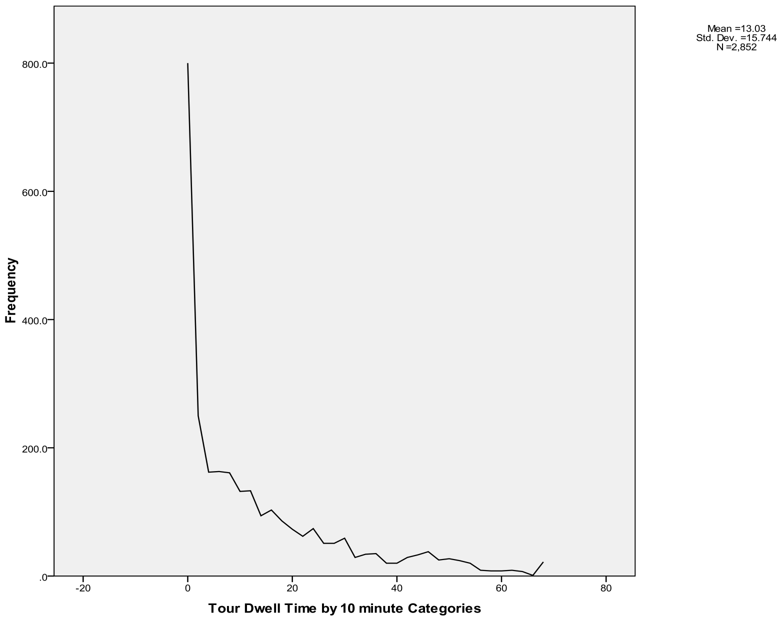

Figure 6: Distribution of Dwell Times within Tours

Figure 7 shows the distribution of the overall duration of the tour. The distribution peaks initially in category 2 which is 20-29 minutes, and then falls, slowly at first and then rapidly. There is a small peak at around category 50, which corresponds to 8 hours and 10 minutes to 8 hours and 20 minutes and would encompass the working day for most people that work. There are rather few full-time workers in this data set, or the peak would be more pronounced in this area of the graph. The slight upturn at the end of the graph is due to designating the highest category as being in excess of 12 hours and 30 minutes. If the categories were continued to higher values, a continual decline in the graph would be found.

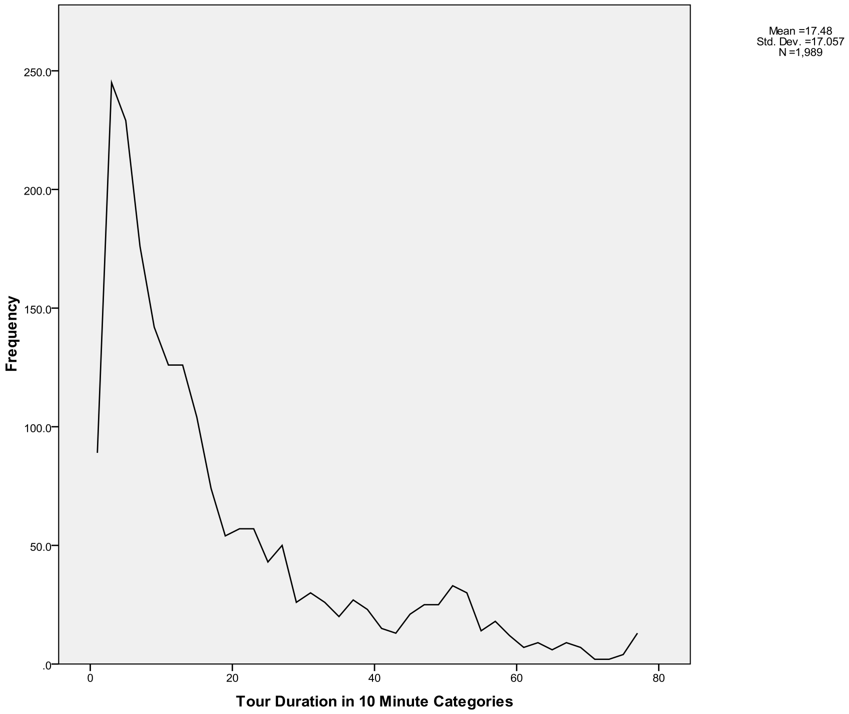

Figure 7: Distribution of Total Duration of Tour

The distribution of start times of tours is quite interesting and somewhat different from the start time of trips as shown in Figure 8. The distribution of tour start times shows that the majority of tours start in the morning and there is a steady decline in tour starts as the day progresses, although there is a small peak in the late morning and another in the early afternoon. Very few tours start after 7 p.m. (19 hours).

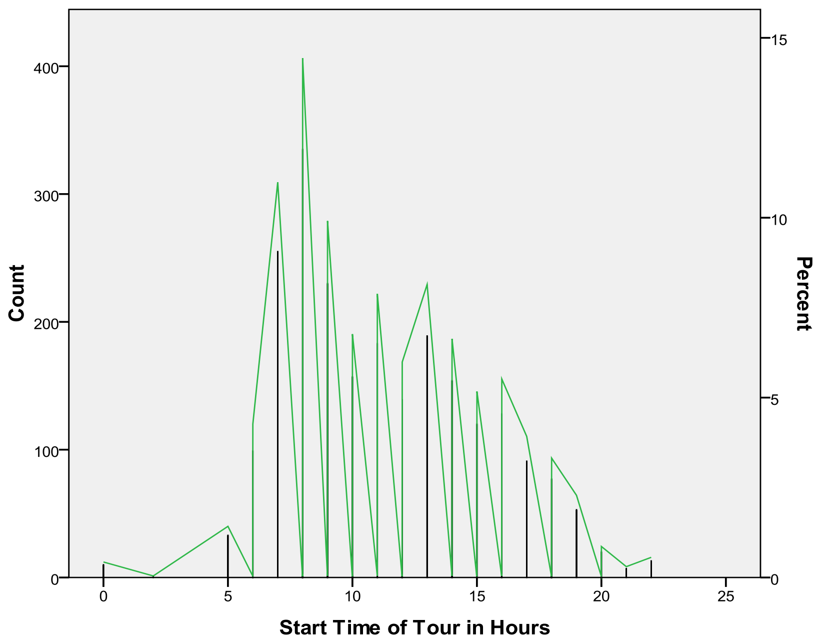

Figure 8: Distribution of Tour Start Times

Figure 9 shows when the tours end, and shows almost the mirror image of Figure 8, as would be expected. In this case, few tours end before 9 a.m., but the peak of tour ends is at about 3 p.m. (15 hours) and is still high at 4 p.m. There is a rapid decline after 6 p.m. (18 hours). Again, if there were a larger proportion of workers in the data set, we would expect to see a peak somewhat later in the afternoon, and the peak that does show here is probably associated more with school ending and the end of tours initiated in the morning by nonworking people.

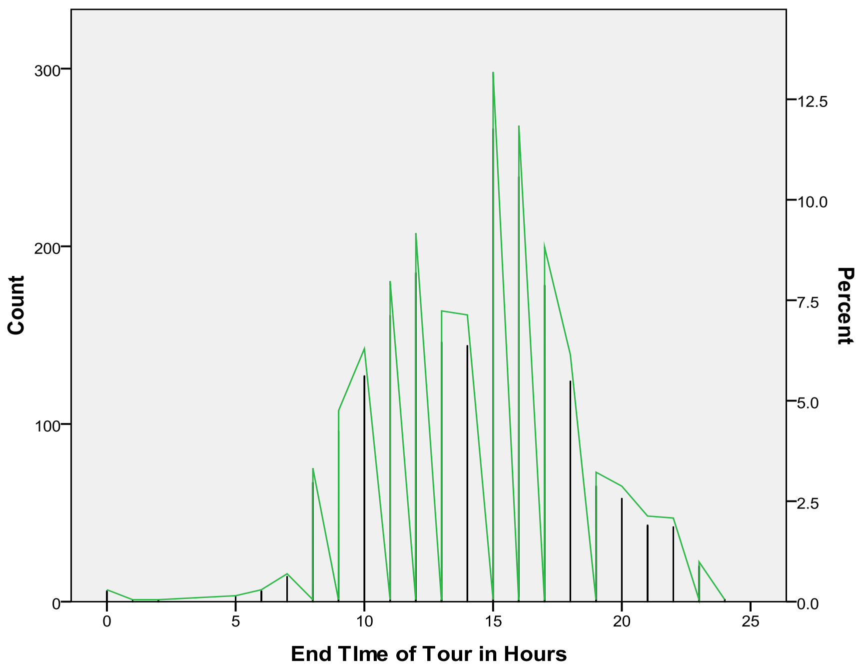

Figure 9: Distribution of Tour End Times

Many other statistics can be reviewed for the tours and trips from the data set. However, these statistics and distributions serve to indicate the nature of the data available and also show that the travel patterns, at least as revealed from the trip data, are much as one would expect.

SUGGESTED CLASSIFICATION OF TOURS BY PURPOSES AND COMPLEXITY

A number of researchers have put forward potential tour definitions and many of these vary quite significantly from one another. There also tends to be some confusion of both definition and treatment of tours and trip chains. The authors of this paper prefer to use the definition of tours and chains put forward by O'Fallon and Sullivan (2005). In their research O'Fallon and Sullivan defined a tour as a sequence of trips and activities that begin from home and return to home. A trip chain, however, is a sequence of trips and activities that begins from a point where a person has spent at least 90 minutes and continues until an activity location where the person next spends at least 90 minutes. Thus, a trip chain could sometimes be the same as a tour, if the only location where the traveller spends at least 90 minutes is home, irrespective of the number of stops along the chain. In other cases, a tour may consist of multiple chains, especially if work or educational purposes are included in the stops within the tour.

Using as a starting point the work of O'Fallon and Sullivan (2005, 2009), a set of twelve tour classifications are put forward, based on a hierarchy of trip purposes of work, education, shopping, and other. The mutually exclusive and exhaustive set of possibilities are shown in Table 1 for the twelve classes of tour.

In Table 1, the letter 'h' stands for home, 'w' for work, 'e' for education, 's' for shopping, and 'o' for other. Also, in Table 1, the square bracketed trip purposes must occur in the sequence, with the bold purposes occurring at least once for each bolded purpose in the sequence. The purposes in round brackets may not occur or may occur multiple times within the sequence. For example, the sequence h – [**w**/o] – (– w/o –) – [**w**/o] – h includes h – w – w – h, h – w – o – h, h – o – w – h, h – o – w – o – w – o – h, etc. Indeed, this sequence includes all possible permutations of o and w between two hs, does not include any e or s purposes, must include at least three trips, and must begin and end at home.

Tour classes 1 through 4 are simple tours, involving two trips and one non-home destination. Tour classes 5 through 7 are complex tours that include at least two stops and three trips, but may include stops for other purposes. Tour classes 8 through 10 include at least two stops and three trips, and must include at least two primary purposes (work, education, or shop) and may also include other purposes. Tour class 11 must include at least three stops and four trips, and must include at least one of each of work, education, and shopping, and may also include other purposes. Tour class 12 must include at least two stops and three trips, and must not include any of work, education or shopping purposes. The order of purposes in the sequences is not important and any order of the trip purposes specified in the sequence is permissible.

Table 1: Proposed Tour Type Classifications

Tour Type Number	***Tour Description***	***Sequence***	***Count of Tours in Panel***
1	Simple work tour	h – w – h	61
2	Simple education tour	h – e – h	39
3	Simple shopping tour	h – s – h	148
4	Simple other tour	h – o – h	888
5	Complex work tour (including composite and multi-part work tours)	h – [**w**/o] – (– w/o –) – [**w**/o] – h	67
6	Complex education tour (including composite and multi-part education tours)	h – [**e**/o] – (– e/o –) – [**e**/o] – h	100
7	Complex shopping tour (including composite and multi-part shopping tours)	h – [**s**/o] – (– s/o –) – [**s**/o] – h	427
8	Complex work and education tour	h – [**w/e**/o] – (– w/e/o –) – [**w/e**/o] – h	41
9	Complex education and shopping tour	h – [**e/s**/o] – (– e/s/o –) – [**e/s**/o] – h	23
10	Complex work and shopping tour	h – [**w/s**/o] – (- w/s/o -) – [**w/s**/o] – h	0
11	Complex work, education, and shopping tour	h – [**w/e/s/**o] – [**w/e/s/**o] – (– w/e/s/o –) – [**w/e/s/**o] – h	0
12	Multi-part Other Tour	h – [**o**] – (– o –) – [**o**] – h	480

This is an exhaustive and mutually exclusive classification of tours, using the hierarchical ordering of trip purposes of work, education, shopping, and other. To test the definitions, they were applied to the panel data described in the previous section with the results shown in the last column of Table 1. Again, the lack of workers in the sample is clear from the rather small number of simple work and complex work tours. The lack of tour types 10 and 11 is also likely to be due to the lack of workers in the sample. However, it would be expected that these tour types would occur much more frequently in a sample that had a better representation of workers.

Overall, the authors feel that this classification of tours is useful and workable. It is being used in further explorations of the data towards a new tour-based modelling approach.

ANALYSIS OF TOUR MAKING

To indicate the approach that can be taken with the tour data available from the GPS measurement, this section describes some preliminary work that has been done using the method of Classification and Regression Trees (Brieman et al., 1984). Decision tree analyses were undertaken for weekday travel in the Wave 2 Add-on dataset. CART® (by Salford Systems) software was used for analyses on the target variables of:

- Number of Tours per Day (NumToursPD),
- Number of Trips per Tour (NumTripsPT), and
- Tour Duration (TourLength).

The predictor variables that were tested include:

- Family Life Cycle (FLC),
- Household Size (HHSize),
- Vehicle Ownership (NumVeh),
- Driver Licence (Licence),
- Employment status (isEmployed),
- Study status (isStudy), and
- Gender.

The Family Life Cycle (FLC) is a composite variable that takes into account the family structure, adult partner, children and children's age. Table 2Table shows the coding and description of the FLC variable.

Table 2: Family Life Cycle Variable Code and Descriptions

FLC	**Description**
1	Single person
2	Couple only
3	Couple, youngest child aged under 18 years
4	Couple, youngest child aged 18 years or over
5	Single parent family, youngest children aged under 18 years
6	Single parent family, youngest children aged 18 years or over
0	Other

Results of the Decision Tree Analyses

Analysis of Number of Trips per Tour

The maximum number of trips per tour in the dataset was 31 trips. The tree was grown with all cases in the dataset (total 1,960) and the resulted best performance tree had a sole splitter of Employment. The same result occurred for trees grown with cases up to 12 trips per tour. It suggested that work related travel has a significant effect on the decision trees although there were only 23 cases with 12 or more trips in each tour.

The tree grown with a maximum of 11 trips per tour had all seven predictive variables as splitters on the tree, but employment remained the first primary splitter on the top of the tree.

However, testing with a constraint of a maximum of 10 trips per tour showed family life cycle replacing employment as the first primary splitter.

Figure 4 (above) shows the frequency of the number of trips per tour in the data. The frequency table on which the figure is based shows 84% of the cases in the dataset have 4 trips or less in the tour. Therefore, it is important to grow a tree with a further constraint to the maximum number of trips in a tour to uncover key determinant variables on predicting tour complexity.

The best performing tree with a maximum of four trips per tour constraint has four primary splitters including family life cycle, number of vehicles, gender, and driver's license as shown in Figure 10. The tree details are shown in Figure 11. The tree details reveal that households with children aged younger than 18 years are split to the right at the first level of the tree with gender as the only other splitter on the right branch at the next and last level of the split. The left branch of the tree has two levels of split being:

- Split by number of vehicles at 1.5 vehicles (i.e. 0 and 1 vehicle to the left child node, and 2 or more vehicles to the right child node)
- The right node is further split by the driver's license variable with grouping of Full and Provisional Licence to the left child node, and No Licence and Learner Licence to the right child node.

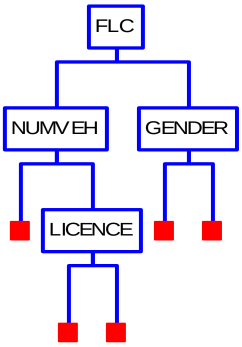

Figure 10: Primary Splitters in the Number of Trips per Tour Tree

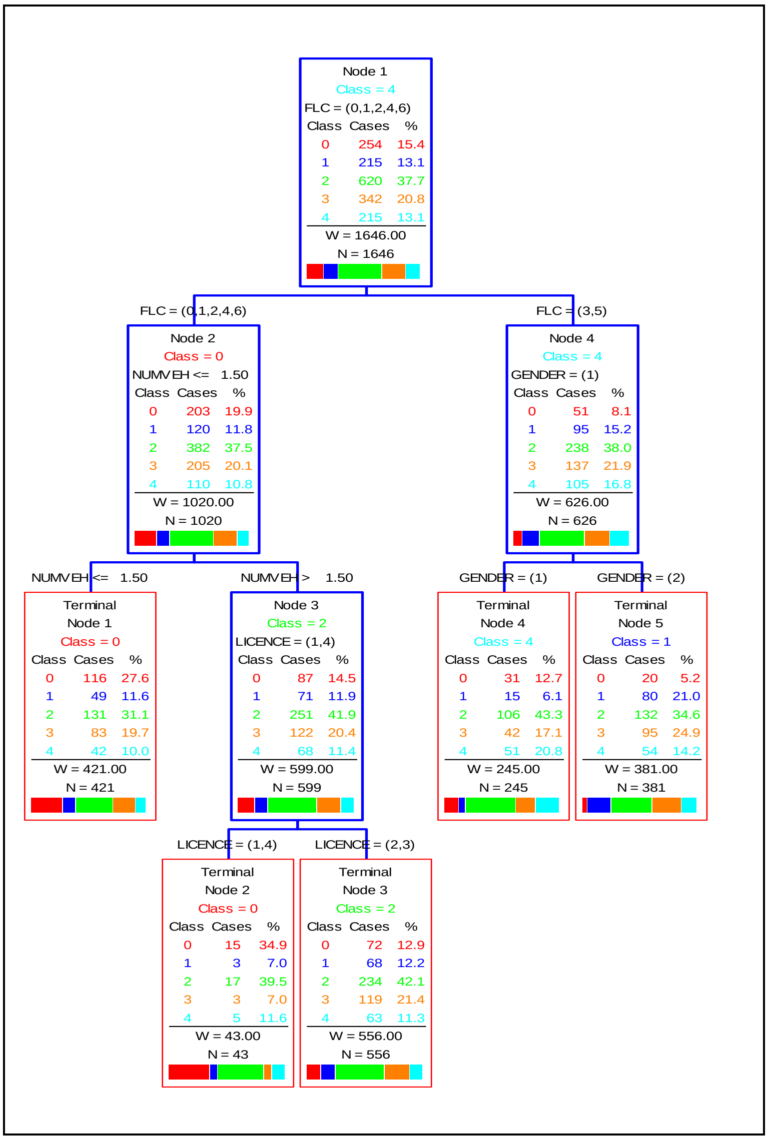

Figure 11: Number of Trips per Tour Tree Details

Number of Tours per Day

In the case of the number of tours per day, Figure 3 shows that few respondents made more than three tours in a day, with the maximum number of tours in a day being eight. For the purposes of the classification analysis, the number of tours per day were limited to three. As with the number of trips per tour, family life cycle was usually the first of the primary splitters in the classification process. It was generally followed by vehicle ownership and driver's license status. Employment status, when it appeared in the classification tree, was about the fifth variable in importance. It is interesting to note that life cycle stage 5 is particularly associated with two and three tours per day, followed by gender (female), suggesting that an important group of multiple tour makers are women who are single parents with children aged under 18 years living at home. While this is not a surprising result, it does indicate that the classification trees are producing useful and sensible results. Again, as in the classification trees for numbers of trips per tour, car ownership generally splits at a value of

1.5, indicating that households with zero cars or one car behave differently from households with 2 or more cars.

Tour Length

The analysis of tour length in this data set is somewhat hampered by two aspects. First, as noted earlier in this paper, the sample for this panel wave had relatively few people employed full time. This tends to reduce the number of tours that would take in excess of eight hours to complete. Second, the GPS device that was being used at this time was one that had a normal battery life of only 8 hours or so, so that it is likely that a number of return trips home at the end of a working day were not recorded, because of loss of battery power. This specific aspect of the data is yet to be fully investigated.

The tour length regression trees split the tour lengths into a number of rather arbitrary groupings, with the shortest tours being under 38 minutes (24 observations), a major category being tours between 38 minutes and 1 hour and 38 minutes (875 observations), and the longest tours being those of over 5 hours and 36 minutes. However, similar to the previous analysis, stage in the family life cycle proved again to be the primary splitting variable in all cases. Here it was found that households in life cycle stage 3 tended to be associated with most of the longer tours. These are couples with children under the age of 18 present in the home. In these cases, the second variable is always the household size, where the longest tours are associated with household sizes greater than 3 and the shorter tours are associated with household sizes of 3 or less. In those branches of the tree where household size appears as the second variable, gender is always the third variable. In all the other cases, employment status is the second variable, with life cycle stage 2 (couple only) appearing as the third variable. Number of vehicles in the household is the fourth primary variable to aid in splitting the tour length, irrespective of which variables preceded it.

Variable Importance

Variable importance is a key measure in decision tree analysis on the contribution of each variable to the tree growth. The higher a variable's importance score, the more contribution the variable has on the splitting of the tree branches. Table 3 summarises the variable importance score of the three trees. It should be noted that the tree for the number of tours per day was constrained to a maximum of three tours per day, while the tree for the Number of trips per tour was constrained to a maximum of four trips per tour. The tree for tour duration is a regression tree whilst the other two trees are classification trees.

The results show that the top three important variables are Family Life Cycle, Household Size, and Employment except in the tour duration tree where employment is replaced by the Number of Vehicles.

Table 3: Variable Importance Scores

	Tree		
Variable	**Number of Tours Per Day**	**Number of Trips per Tour**	**Tour Length**
Family Life Cycle	100.00%	100.00%	100.00%
Household Size	79.49%	75.70%	64.25%
Employment	61.98%	87.35%	19.75%
Driver's License	46.73%	75.30%	19.67%
Number of Vehicles	36.69%	53.46%	85.31%
Gender	27.54%	82.74%	63.84%
Student Status	25.81%	11.71%	11.43%

CONCLUSIONS

First, the research reported in this paper has demonstrated quite clearly the wealth of data available on tours from GPS data. It has shown that it is entirely feasible to develop estimates of tour-making behaviour from GPS data collected over multiple days from a rather small group of individuals. Second, a tour classification scheme has been proposed that clearly differentiates between simple and complex tours, and that also embodies aspects of trip purposes within the tour definitions. It was found that, even in the rather small data set used for this research, the different tour classes were quite well represented. Using these tour classifications in the data has also provided a useful starting point for future modelling exercises that are expected to be the next stage of this research.

Finally, the classification and regression trees analysis of the GPS tour data has shown some interesting results, with the strongest result being that the number of trips in a tour, the number of tours undertaken in a day, and even the total duration of a tour are most strongly related to the stage in the family life cycle of the household from which the tour maker comes. This indicates that this is likely to be found to be an important modelling variable for models of the number of tours per day and the number of trips in a tour. Second to the stage in the family life cycle as an influencing variable are such measures as the household size, the number of vehicles available to the household, and employment status.

The use of GPS data as a basis for improved modelling of travel behaviour appears very promising from this research. Using the tour classification scheme proposed in this paper may also provide a useful method for modelling, compared to many of the current approaches that attempt to model a large number of different patterns of tour making.

REFERENCES

Axhausen, K.W., A. Zimmermann, S. Schonfelder, G. Rindsfuser, T. Haupt (2002).Observing the rhythms of daily life: A six-week travel diary. *Transportation* **29**, 95-124.

Axhausen, K. W., M. Löchl, R. Schlich, T. Buhl and P. Widmer (2007). Fatigue in long duration travel diaries, *Transportation*, Vol. 34, No. 2, pp. 143-160.

Battelino, H. and J. Peachman (2003). The Joys and Tribulations of a Continuous Travel Survey, in *Transport Survey Quality and Innovation*, P.Stopher and P. Jones (editors), Pergamon Press, pp. 49-68.

Bowman, J. and Ben-Akiva, M. (2001). Activity-based disaggregate travel demand model system with activity schedules, Transportation Research Part A: Policy and Practice, 35(1), 1-28

Brieman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. (1984). *Classification and Regression Trees*, Chapman and Hall.

Giaimo, G.; Anderson, R.; Wargelin, L.; Stopher, P. (2010). Will It Work? Pilot Results from the First Large-Scale GPS-Based Household Travel Survey in the United States. Paper presented to the Annual Meeting of the Transportation Research Board, January.

Hanson, S. and J. Huff (1988) Systematic variability in repetitious travel. *Transportation* 15, 111-135.

O'Fallon, C., & Sullivan, C. (2005). *Trip chaining: understanding how New Zealanders link their travel.* Wellington: Transfund New Zealand.

O'Fallon, C.; Sullivan, C. (2009). Trends in Trip Chaining and Tours: Analysing Changes in New Zealanders' Travel Patterns Using the Ongoing NZ Household Travel Survey. Wellington, Pinnacle Research & Policy Ltd, Wellington.

Pas, E.I. (1986) Multiday Samples, Parameter Estimation Precision and Data Collection Costs for Least Squares Regression Trip-Generation Models. *Environment and Planning A*, 18, 73-87.

Stopher, P., N. Swann, and T. Bertoia (2006). Trip Rates, Non-Mobility, and Response Rate: Measures to Evaluate the Quality of a Travel Survey, paper presented to the 29th Annual Meeting of the Australasian Transport Research Forum, Brisbane, September.

Stopher, P.R.; Clifford, E.; Zhang, J.; FitzGerald, C. (2008a). Deducing Mode and Purpose from GPS Data. Paper presented to the 87th Annual Meeting of the Transportation Research Board, Washington, D.C., January.

Stopher, P.R.; Zhang, Y.; Halling, B. (2009). Results of a TravelSmart Evaluation in South Australia. Paper presented to the 30th ATRF Meeting, Auckland, New Zealand, September.

Stopher, P.; Wargelin, L. (2010). Conducting a Household Travel Survey with GPS: Reports on a Pilot Study. Paper submitted to the WCTR Conference, Lisbon, July.

Stopher, P.R.; Prasad, C.; Zhang, J. (2010). Comparing GPS and Prompted Recall Data Records. Paper submitted to the WCTR Conference, Lisbon, July.

USDOT (2008). Ntional Household Travel Survey. http://nhts.ornl.gov/index.shtml Accessed 6/01/2010.

Vovsha, P., M. Bradley and J. L. Bowman (2005), Activity-Based Travel Forecasting Models in the United States: Progress Since 1995 and Prospects for the Future, Progress in Activity-Based Analysis, Elsevier.