Exploring the Full Continuum of Travel:
Data Fusion By Recursive Partitioning Regression

Heather Contrino
Bureau of Transportation Statistics

Nancy McGuckin
Travel Behavior Analyst

David Banks
Bureau of Transportation Statistics

EXPLORING THE FULL CONTINUUM OF TRAVEL: DATA FUSION BY RECURSIVE PARTIONING REGRESSION

ABSTRACT

The analyses that can be conducted in a given public policy area are often limited by the content of a single database but often require data that are available from more than one source (Rodgers 1984). It is useful if data from multiple databases can be combined (fused) to support planning and decision-making. Except in the simplest cases, such fusion poses hard statistical problems. This paper describes an effort to address some of these problems through the use of recursive partitioning regression (CART).

To illustrate our approach, we take the first steps to fuse household travel data that were obtained from two surveys performed for the United States Department of Transportation in 1995. The first data set is from the Nationwide Personal Transportation Survey (NPTS); this survey was performed for the Federal Highways Administration, and focused upon daily travel (mostly short trips). The second data set is from the American Travel Survey (ATS), which was run by the Bureau of Transportation Statistics (BTS); the ATS focused on long-distance travel (trips of 100 miles or more). Both surveys collected data on household demographics and trip characteristics from the same population, but the surveys did not include the same households in the sample. Clearly, a more complete understanding of all passenger travel in the U.S. will be obtained if households in the surveys can be meaningfully matched, so that long and short trips may be studied together.

This paper describes the need for data fusion in federal statistics and outlines previous work on data fusion approaches. We detail the recursive partitioning analyses we perform, and interpret the results. Finally we address the problem of data fusion using the refined partition, and place this in the context of statistical methods for record matching. The last section draws conclusions about the value of the proposed methodology.

1. OVERVIEW AND PURPOSE

In the United States, the federal government maintains thousands of data collection programs, each focused around some specific program goal. But with rising survey costs and the need to improve the efficiency of government spending, there is great pressure on government statisticians to use these data sets to address questions that were not foreseen during survey design.

The experience of the United States Department of Transportation (DOT) is typical. Large amounts of data are needed by many different agencies within DOT to track program performance and assess new needs. These agencies undertake separate collection efforts, but portions of those efforts, especially questions related to demographics, are very similar. So when transportation professionals seek to look across programs or develop general information on transportation behavior, it is natural to

attempt to combine the information from databases maintained by the different agencies. Additional impetus for data fusion derives from the Paperwork Reduction Act, which strongly discourages respondent burden and thus implicitly encourages the division of lengthy questionnaires into parts that can be distributed across different households. This forces a compromise on the completeness, quality, and utility of data that are collected.

The collection of passenger travel data is especially troubled by this compromise, because travel information is so complex. Besides necessary household demographics, detailed trip data, such as the mode, purpose, origin, destination, distance, and so forth of each trip are obtained. The required detail imposes significant burden, and consequently issues of low response rates, curtailed interviews, and suspect quality are of top priority. The current approach to addressing this is to use multiple instruments that collect information on separate aspects of travel (Madre 1999).

For example, the U.S. DOT currently has two national surveys in place that focus on different types of trips and use different trip definitions and reporting periods:

- The Nationwide Personal Transportation Survey (NPTS) collects data on all trips made by U.S. households during one day. Although the survey collects data on trips of all lengths, because of the short reporting period, the emphasis is primarily on daily, local travel.
- The American Travel Survey (ATS) collects data on trips of 100 miles or more (one-way). Because of the rarity of these trips, a three-month reporting period window was used.

In 1995 both the NPTS and the ATS were conducted. In this paper we look at the possibility of fusing these two datasets to create one complete representation of the full continuum of personal travel by U.S. households.

These two national surveys were conducted separately, with no coordination in the sampling, data collection methods, and questionnaire design. Both surveys provide valuable information; however, these historically uncoordinated efforts leave large gaps in passenger travel data. The NPTS provides an abundance of data on short, local trips while the ATS does not even ask for this information. This is an unfortunate disconnect; it means that the DOT has little data on the full continuum of passenger travel from which to guide planning and policy-making at a national level.

A second concern is that the separate DOT surveys provide no useful data on the relationship between long-distance and daily travel within the same households. There is great interest in improving our understanding and ability to predict travel behavior. However, one factor affecting daily trip rates may be the rate of long-distance travel in the household. Without data on all travel, an important piece of the picture may be missing.

A third consideration is our ability to check the completeness and accuracy of the collected data. Jean-Loup Madre (1999), in the context of the French National Travel

Survey, recommends a joint daily and long-distance survey, in part because of the benefits derived from crosschecking the data.

Together, these data sets could provide more precise and thorough information on the characteristics, behaviors, and preferences of transportation users so that the performance of the transportation system can be accurately measured and improved based on current and future needs. Because of the differing reporting periods and the amount of information needed on each trip to establish seasonal trip rates by mode, purpose, and length for regions and population subgroups, collecting data on both daily and long distance travel from each sampled household may be too burdensome for respondents. In light of this, we need an alternative method for obtaining data on the full continuum of travel in the United States and the relationship, if any, between long-distance and daily travel.

One such alternative is to make use of the data fusion techniques to link household data on daily and long distance travel (Rassler and Feischer 1997, Kamakura and Wedel 1996). The rest of this paper examines issues and methods for achieving such linkage using NPTS and ATS data as the motivating example and testbed.

2. OVERVIEW OF FUSION PROCESS

Data fusion involves the linking of two survey data files from the same population based on a set of common variables. Mostly, the fusion of data is applied when an alternative single source data set cannot be sampled for practical reasons (Wiedenbeck 1999). The objective is to expand the scope of information available in any given source by matching records from one source to a second source using a set of common variables and matching criteria (Rodgers 1984). Statistical matching procedures were developed by analogy with exact matching procedures, where records from one source are linked to records from a second source using unique identifiers (Rodgers 1984).

When data comes from different samples it is not possible to do an exact match on the same individual. Therefore a fusion process has been developed to allow the matching of information from each survey based on a set of common variables. The process, in theory, is very close to the imputation process applied to missing data in one data set in that missing data is filled in with data from a record or grouping of records most like the missing record. In fact, the data to be transferred or appended is often called the "missing data" (Baker, Harris, and O'Brien 1994).

In data fusion, information from one survey is appended to another survey based on common variables of the sample units to enhance the scope of information available for analyses. This is carried out at the individual record level. Each record in one of the data sources (donor file) is matched with a record from the second source (recipient file) (Rodgers 1984). The accuracy of the data fusion process is a function of the donor file; the larger the donor fine, the better chance there is of finding more acceptable matches between donor and recipients (Baker, Harris, and O'Brien 1989, Rassler and Feischer 1997).

The objective of data fusion can be viewed as the analysis of an unknown common distribution [X, Y] of two multivariate variables X and Y (Wiedenbeck 1999). Since there is no single source with information on X and Y together, an artificial data base (Z) is generated by matching the observations of both sources according to common variables (Rassler and Feischer 1997). For example, let X be the recipient file and let Y be the donor file. Individual records from file Y are appended to individual records in file X to form a new data set Z that contains data on X and Y.

X= recipient file
Y= donor file
Z= X+Y, based on the matching of variables common to X and Y.

Each recipient record is linked to one or more donor records in file Y, on the basis of variables that are observed in both files (Kamakura and Wedel 1996). The key principle of fusion is that once respondents from the donor file and recipient files are matched in some way, all the missing data are passed from donor to recipient, thus preserving the interrelationships between variables from the donor to recipient file (Baker, Harris, and O'Brien 1994). Rodgers (1984) provides an in-depth description of the various procedures developed for data fusion.

In the case of the NPTS and ATS, the objective is to append the long-distance trip data (ATS) onto the daily trip data (NPTS) to allow for analyses of the full continuum of passenger travel and the relationship of long-distance and daily travel.

2.1 The Common Variables

The selection of common variables is an important part of the fusion process. Typically, several sociodemographic variables are the basis for the set of matching variables primarily because these are the variables most likely to appear in both data sets (Rassler and Fleischer 1997, Kamakura and Wedel 1996). However, many other variables may be included in the files for the specific purpose of data fusion (Kamakura and Wedel 1996).

Various researchers take slightly different approaches, but typically the common variables are determined by multiple regression where the variables with the highest R^2 squared in the donor and recipient files are used for the matching process. The individual records are then matched based on either an exact variable match (e.g. sex) or by use of a distance function (e.g. similarities in age).

Baker, Harris, and O'Brian (1994) devote a significant portion of their experiment to choosing common variables and matching criteria. They use discriminate analysis and regression analysis to find the predictive common variables and suggest that an analysis of the donor file be carried out prior to the matching process in which each variable to be fused is cross-tabulated against all potential common variables and included if discrimination is evident. O'Brien (1991) on the other hand, uses all variables common to both data sets but assigns different weights to each variable based on the results of the analyses of variance. Kamakura and Wedel propose a probabilistic model that first

identifies homogenous groups on the basis of all information available from the two samples as the first step in matching variable selection (1996).

Much of the research, actually, provides little information on the process of selecting matching variables. In contrast, researchers have focused more extensively on the matching process and the validity of various matching algorithms (Rubin 1986, Rodgers 1984, Wiedenbeck 1999, Rassler and Feischer 1997). Rassler and Fleischer (1997), for example, provide a thorough overview of the matching process and offer an in-depth examination of the performance of various fusion algorithms. However, little emphasis is placed on the determination of matching variables.

We felt that the selection of match variables deserved greater attention, and therefore aimed much of our focus on that process. This focus on selecting the right set of variables led us to the use of recursive partitioning regression. This methodology is widely associated with the acronym CART (for Classification and Regression Trees) as popularized by Breinam, Friedman, Olshen and Stone, 1984.

To achieve the household matching, we use CART to develop recursive partitioning regression models for both data sets, where the number of trips in each household is taken as the response variable. These CART models divide the space of the explanatory variables into regions such that the variation in trip count for households within the same region is relatively small. The two analyses thus produce two independent partitions of the space of explanatory variables. We argue that if these partitions are overlaid to produce a more refined partition, then households within the sub regions induced by the refined partition may be matched for certain kinds of statistical analyses**.**

This approach allows us to use what we know about travel behavior and the predictive demographics, and explore how these demographic variables could be used to bridge the gap between the two datasets. To make a meaningful match, we make an in-depth examination into the selection of matching variables and provide an alternative to logistic regression for the selection of matching variables: recursive partitioning regression.

3. RECURSIVE PARTITIONING REGRESSION

Recursive Partitioning Regression (RPR) is a computer-intensive competitor to conventional multiple linear regression. It is designed to work well when the functional relationship is nonlinear, when certain explanatory variables are only applicable in certain regions of the variable space, and when variable selection is required. Although an early version of RPR was proposed by Morgan and Sonquist (1963), the method did not become popular until computational speed and more sophisticated statistical theory led to CART, a software program described by Breiman, Friedman, Olshen and Stone (1984). CART is an acronym for Classification and Regression Trees; most often it is used in classification problems. In this paper we focus on its performance as a tool for regression.

In recent years, many implementations of RPR have been developed---we used the

version of CART that is commercially available from Salford Systems, Inc., but our usage should not be interpreted as any recommendation of that product. Different implementations effect slightly different strategies for deciding when partitions provide useful explanation. These choices can lead to somewhat different regression trees, and at present there is no clear consensus within the statistical community as to which approaches work best under any given circumstances. We are generally pleased with our results, but recognize that a comparative study would be worthwhile. (For further discussion of the robustness of the CART results, see Shannon and Banks, 1997)

An RPR algorithm works by considering all possible splits on all possible explanatory variables. For continuous variables, such as age, it examines all the gaps (here age was recorded to the nearest adult year, so it could split between 18 and 19, 19 and 20, and so forth); for categorical values, such as whether or not one is a driver, it considers each category. RPR picks as the first split the one that does the best job of separating low response values from high response values. Then it reapplies the same procedure recursively, first to the cases on the left-hand side of the initial split, and then to the cases on the right. Proceeding in this fashion, it generates a regression tree. When it reaches a point in a branch of the tree at which no explanatory variable splits the cases sufficiently well to satisfy the implementation's splitting criterion, then the algorithm stops splitting and declares a terminal node. Different RPR algorithms employ different kinds of splitting and termination criteria; additionally, some RPR algorithms, such as CART, grow large trees, and then prune them back using cross-validation to reduce prediction error.

In this study, we used:

- The least squares criterion to divide the cases at splitting nodes. The least squares rule picks the split that most reduces the mean squared error in the two new groups, as compared to the mean squared error in the original undivided group.
- A complexity penalty that prevented subdivision of nodes with fewer than ten cases. The complexity penalty and tree pruning ensure that trees do not have too many branches. (For the ATS, the optimal tree that was produced had 40 terminal nodes---although this gave good fit, we believed that for this initial study of RPR methods in data fusion, it was preferable to increase the complexity penalty, resulting in a tree with only 20 terminal nodes.)
- 10-fold cross-validation to improve prediction error.

Cross-validation addresses a subtler problem. This technique is used to prevent overfit. In applications with many variables, it is easy to find spurious regression structure by chance; this leads to good fit for the training sample, but poor predictions with future data. Essentially, one is fitting the noise in the sample as well as the signal. A common way to control this is to divide the sample at random into ten parts; for each part, an RPR tree is built with the remaining nine-tenths of the data, and then used to predict the held-out tenth. The final tree is pruned back until the prediction error on the hold-out subsamples is minimized.

The following characterizations are, very slightly, impressionistic; for example, we refer

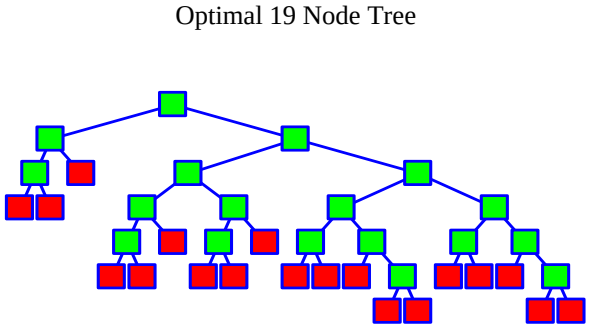
to urban and non-urban areas, but realize this is not a sharp distinction. Similarly, we refer to two-parent households, but recognize that some of these might contain a grandparent, parent, and child. But for broad descriptive purposes, we try to match conventional household patterns to the terminal nodes in the CART tree.

With this perspective (more details on the algorithm are contained in Breiman et al., 1984), we proceed to interpret the CART trees produced in the study of the NPTS data. The quickest way to describe the RPR method is to work though an example of CART output from the analysis of NPTS data. In this analysis we took the number of trips to be the response variable, and built a model that predicted that response from the values of the explanatory variables, which were essentially all of the other variables collected in the survey.

3.1 Description of the Variables Effecting Daily Travel

Figure 1 shows the regression tree that CART built.

Figure 1: NPTS CART® Tree



To discuss Figure 1, note that the tree contains two types of nodes: terminal nodes and splitting nodes that simply divide the dataset. The terminal nodes show the estimated value (in this case the trip rate) for the specific combination of response variables when no other splits in the database are meaningful. The splitting nodes are where the program separates the households in the survey according to some value of an explanatory variable. To facilitate description of the tree, we number the splitting nodes from top to bottom and from left to right. Similarly, the terminal nodes are given alphabetic labels, from left to right.

Table 2a – Definition of Splitting Nodes

NODE	SPLITTER	SPLITTING CRITERION	STANDARD DEVIATION	AVERAGE TRIPS	N
1	DRIVER	No	3.150	4.414	33763
2	AGE	Less or equal to 72.5	2.589	2.587	3295
3	WORKER	No	2.630	2.93	2454
4	LIFE CYCLE	Adult(s), No Children; >1 Adult Child 16-21; Adult(s), Retired No Children	3.142	3.142	30468
5	EDUCATION	Less or equal to high school	2.992	4.360	18791
6	WORKER	No	2.962	4.062	8150
7	EDUCATION	Less then high school	3.049	3.742	3739
8	AGE	Less or equal to 76.0	2.995	4.558	10641
9	LOCATION	New England, Middle Atlantic, S. Atlantic, W.S. Central, Pacific, Chicago, New York	2.996	4.625	10187
10	EDUCATION	Less or equal to high school	3.33	5.017	11677
11	WORKER	No	3.257	4.670	4709
12	LOCATION	Middle Atlantic, E.N. Central, E.S. Central, W.S. Central, Chicago, Los Angeles, New York, San Francisco, Washington DC	3.437	4.267	1066
13	SEX	Male	3.193	4.788	3643
14	RACE	Black, Asian, Other	3.287	5.076	1661
15	HHSIZE	Less or equal to 3.5	3.359	5.252	6968
16	LIFE CYCLE	>1 Adult, Child 0-5	3.164	4.912	2079
17	RELATIONSHIP*	Parent, Sibling, Other, Partner, Non-relative	3.428	5.396	4889
18	LOCATION	Middle Atlantic, S. Atlantic, E.S. Central, pacific, Chicago, Los Angeles, New York, San Francisco, Washington DC	3.429	5.437	4735

* Relationship with respect to the reference person.

The top node (Node 1) in the tree is a splitting node. As shown in Table 2a, it divides the cases in the survey according to whether or not the respondent is a driver. If the respondent is not a driver, Table 2a shows that the next split (Node 2) occurs on age; if the respondent is less than 72.5 years old, two branches occur—one branch goes to terminal Node C (shown in Table 2b)---this node consists of respondents who do not drive and who are more than 72.5 years old. The mean number of trips in Node C is 1.586, and the standard deviation is 2.174. In contrast, the other branch splits at Node 3 according to whether or not the respondent is employed. Those that are employed terminate at Node B, and make an average of 3.404 daily trips with a standard deviation of 2.523. Those that are not employed go left to Node A, and make an average of 2.617 trips, with a standard deviation of 2.652. In this fashion one can interpret the entire tree

in Figure 1, using the information in Tables 2a and 2b.

Table 2b

Terminal Node	Standard Deviation	Average	N
A	2.652	2.617	1479
B	2.523	3.405	975
C	2.174	1.586	841
D	2.890	3.224	1056
E	3.086	3.945	2683
F	2.857	4.333	4411
G	2.959	4.558	8268
H	3.134	4.910	1919
I	2.837	3.775	454
J	3.336	3.951	647
K	3.533	4.757	419
L	3.091	4.546	1982
M	2.972	4.223	274
N	3.319	5.245	1387
O	2.957	4.516	890
P	3.279	5.209	1189
Q	3.140	4.136	154
R	3.334	5.226	2477
S	3.516	5.669	2258

All of the above categories came from the first split of the database on whether the respondent drives. If the respondent is not a driver, the sample divides along age—the elderly from other adults, and then whether the respondent is a worker. If the respondent is a driver (the right hand side of the first split) the next variable of importance is life cycle, or life-stage (Node 4 in Table 2a). The two major divisions of the sample were between respondents in households with young children (less than 16), and respondents in households with older children living at home or no children.

For respondents without young children in the household, the next sample split was on education—high school graduates or less on one side and some college or more on the other. The terminal Node D describes unemployed drivers without childcare responsibilities and less than a high-school education. Node E describes unemployed drivers without childcare responsibilities who have high-school degrees. In Node F are employed drivers without childcare responsibilities and who have high-school degrees. Examination of the average trip rates indicates that employment increases the number of daily trips, as does education, and both effects are plausible. Node G contains non-elderly drivers without child-care responsibilities, more than high-school education, and who live in urbanized areas. Node H is as Node G, except the households are in less urban areas; Node H has a higher trip rate. Node I consists of elderly drivers with more than high-school educations and no childcare duties---they take fewer daily trips than Nodes H and G.

In the other partition (on the right-hand side of Figure 1) of the separation by life cycle (Node 4) are those respondents who have children present in the household. Again the next split divides the data set into those who have a high school education or less: Node J are unemployed drivers with child-care duties and no college education who live in urban areas; Node K are the same, but live in less urban areas. As noted before, those in less urban areas take more daily trips. Node L describes male working drivers with children and no college education. Node M describes minority female working drivers with children and no college education; respondents in Node N are similar but white. The minority women take fewer daily trips than males who take fewer trips than white women (this is an interactive effect between race and gender which could not be easily discovered by conventional multiple regression techniques).

If the respondent went beyond high school, the next variable that divides the data is household size. Node O are drivers in a two-parent household with one child and at least some college. Node P are single-parent drivers with some college education. The latter report more daily trips---this may show that the household sustaining trips that are commonly shared by two parents fall more heavily on single parents.

The next splitting node divides the remaining cases by the respondent's relationship to the head of the household (remember, these are larger households and extended families). If the respondent is not the head of the household, spouse or child, but a parent, sibling, or non-relative then the branch goes to the left. Node Q describes drivers in larger households that have some college but who are part of an extended family—not the spouse or child of the head of the household.

Node R describes educated drivers who are the head of the household in these larger households, or the spouse or child, and who live in an urban environment. Node S is like Node R, except that the environment is less urban.

In looking back over the complex statistical inference described by Figure 1, we note that the variables used for splitting, and the locations of the splits, are plausible. Among the splits, there is some interesting material---the early division of drivers is not surprising, but the large impact of children on daily travel, as shown in Node 4, is good to confirm. . The division based upon high-school education at Node 7 is intriguing, as is the division on race in Node 14. The splits on geography, in Nodes 9, 12, and 18 point up an area of interest to transportation planners, and may ultimately inform efforts to establish more livable communities.

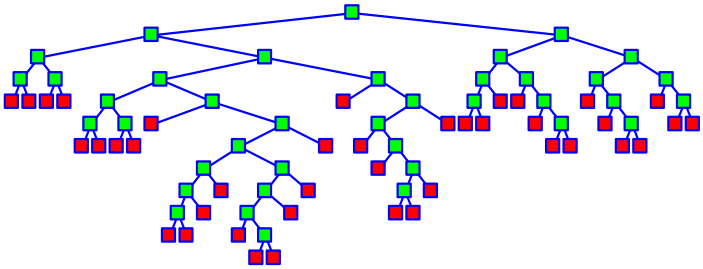
3.2 Description of the Variables Effecting Long-distance Travel

We now contrast the NPTS tree with the ATS tree shown in Figure 2. The ATS tree is slightly more complicated, with 20 terminal nodes, though this represents a significant simplification over the 40 nodes in the optimal tree (here "optimality" refers to the best solution available for our particular choice of tuning parameters in the CART algorithm). It is not surprising that more complexity is needed; the factors that affect short trips in the

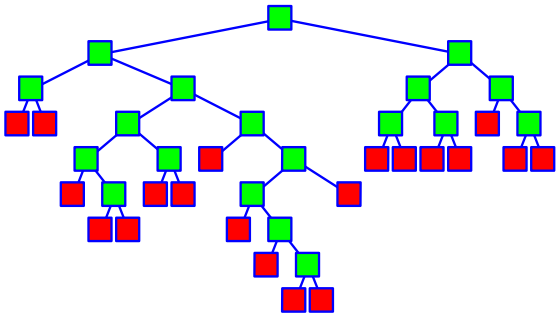
NPTS are likely to be different and fewer than those affecting long-distance travel.

Figure 2: ATS CART® Trees

Optimal 40 Node Tree



20 Node Tree



We read Figure 2 similarly to Figure 1. The first split is on education; those without a college degree go to the left. Node 2 splits on household income, with poorer families moving to the left. Node 3 splits on whether the household owns one or fewer automobiles. Thus cases that terminate at Node A tend to have less education, lower income, and fewer cars than cases in other nodes; as one expects, they also take few long-distance trips. Node B is like Node A, except that they own more cars and take more trips.

Terminal Node C consists of middle-income people who did not attend college and who live in urban areas. Nodes D and E are similar, but have higher incomes---the split that distinguishes these nodes is on employment, and may be spurious, since Node E has only three cases, a very high number of trips, and a high variance. We suspect an outlier is driving the formation of this branch of the tree.

Table 3a

NODE	SPLITTER	SPLITTING CRITERION	STANDARD DEVIATION	AVERAGE TRIPS	N
1	EDUCATION	Less or equal to a bachelor's degree	8.885	5.294	50767
2	HOUSEHOLD INCOME	Less or equal to $31,250	7.833	4.120	38616
3	VEHICLES	Less or equal to 1.5	5.46	2.753	15466
4	LOCATION	New Eng, Mid Atlantic, E.N. Central, S. Atlantic, E.S. Central, W.S. Central, Pacific, Chicago, LA, New York, San Fran, DC	8.964	5.034	23150
5	EDUCATION	Less or equal to high school	8.053	4.368	16977
6	PERSONAL INCOME	Less or equal to $77,500	6.997	3.569	10561
7	WORK STATUS	Working ft, pt, looking for work, Armed Forces, Homemaker, School, Something Else	66.167	24.313	16
8	AGE	Less or equal to 32.5	9.392	5.684	6416
9	MARITAL STATUS	Widowed, Never Married	10.883	6.864	6173
10	AGE	Less or equal to 68.5	11.571	7.412	4919
11	SEX	Female	11.921	7.712	4438
12	HHTYPE	Married-couple family hh, Female family hh, Male hhs with children under 6 only & with no children under 18, Male and female nonfamily hhs	14.985	8.638	2097
13	WORK STATUS	Working full-time, part-time, Armed Forces, Homemaker, School, Retired, Something Else	32.857	20.593	27
14	HOUSEHOLD INCOME	Less or equal to $77,500	10.797	9.023	12151
15	WORK STATUS	Working part-time, looking for work, Armed Forces, Homemaker, School, Retired, Something Else	10.049	8.245	9313
16	AGE	Less or equal to 78.5	8.153	6.640	3362
17	LOCATION	New England, Middle Atlantic, Pacific, Chicago, Los Angeles, New York, San Francisco	10.873	9.151	5951
18	LOCATION	New England, Middle Atlantic, Pacific, Chicago, Los Angeles, New York, San Francisco	12.619	11.577	2838
19	SEX	Female	14.016	13.456	866

Table 3b

Terminal Node	Standard Deviation	Average	N
A	4.127	1.998	8248
B	6.559	3.615	7218
C	6.460	3.537	10545
D	6.952	6.231	13
E	124.837	102.667	3
F	7.469	4.039	1715
G	9.934	6.284	4701
H	7.211	4.715	1254
I	8.172	6.883	2341
J	14.543	8.482	2070
K	21.507	15.654	26
L	0	149	1
M	7.044	4.640	481
N	8.372	7.093	3036
O	3.589	2.417	326
P	9.606	7.460	1936
Q	11.343	9.967	4015
R	10.202	9.276	1276
S	11.431	10.966	696
T	15.502	15.458	866

Node F consists of young people who attended but did not graduate from college and who live in urban areas. They take fewer long-distance trips than their older counterparts in Node G---this may reflect increasing rates of business travel as one moves up in an organization. Node H consists of unmarried middle-income people with some college who live in less urban areas.

Marriage seems to increase the average number of long-distance trips (except for people older than 68.5 years, shown in Node M), but introduces more complexity. For example, Node I consists of married or divorced females with some college and middle incomes who live in less urban areas. Nodes J, K, and L consist of comparable men, with a split at Node 12 that reflects whether they are primary custodians of children (note that Node 12 is a categorical split, some of whose categories are precluded by higher divisions in the tree). Node L has a single case, which should probably be deleted as an outlier; the effect of this deletion could simplify the local structure of the tree.

To the right of the chart (cases that go right at Node 1) are college graduates. Terminal Node N consists of unemployed college graduates; Node O consists of retired college graduates. The latter take far fewer long-distance trips. Node P consists of working middle-income college graduates who live in urban areas; their counterparts in less urban areas, Node Q, take more long-distance trips. Node R consists of upper-income, urban college graduates. Node S are upper-income female college graduates who live in less urban areas; Node T is similar, but consists of males. The men take more long-distance trips than the women.

This description of the CART output is slightly impressionistic, since full specificity would require much more space. But readers should refer to Section 6, which outlines the definitions and categories used.

We are encouraged that, like the NPTS result, the ATS tree picks out plausible variables and splits. The effects of two outliers appear in Nodes E and L, and this could be treated either by removal or use of a robust splitting criterion. But the structure of the tree is sensible. Education plays a very large role, as does employment and income (which are closely related). Location is a recurrent split, with the consistent effect that those who live in urban areas take fewer long-distance trips. Gender appears twice; when it is relevant, women appear to take fewer long trips than men. These findings should be considered in the context of previous travel research, and additional RPR work should be done to test these apparent patterns.

4. COMPARISON OF THE RESULTS TO REGRESSION

Having described the RPR methodology and the trees that CART generated, we now want to compare our results against conventional multiple regression output. Multiple regression is a natural competitor in this arena, and is widely used in data fusion research to identify variables upon which cases should be matched (Baker, Harris, and O'Brien 1994). To motivate the distinction between RPR and regression, imagine we have a single explanatory variable. In that case, regression fits a line, whereas CART fits a staircase function. The former method works best when the true relationship is linear, but the latter is better when the relationship is more complex.

To compare the CART trees with regression, we use mean squared error from the fitted model. In regression, mean squared error is just the average of the squared deviations between each observation and the fitted regression surface. In RPR, mean squared error is the average of the squared deviations between each observation and the average value in its terminal node. The more variables a method uses, the smaller the mean squared error will be (but if too many variables are used, one has overfit, which leads to unreliable inferences).

To compare RPR and regression on the same footing, it is essential to ensure that both use the same number of explanatory variables. The CART analysis of the ATS data used ten explanatory variables and achieved a mean squared error of 69.69. The corresponding regression analysis, using backwards elimination to determine the best model with exactly ten explanatory variables, had a mean squared error of 20.65. Although this indicates that regression obtains better fit, one should bear in mind that regression could only use the 425 cases without critical missing data, whereas CART fit all 50,767 cases. Given this limitation, it is premature to conclude that regression gives better model fit than RPR for ATS data.

Similarly, for the NPTS data, CART used eleven variables to achieve a mean squared error of 9.25 on 33,763 cases, whereas the corresponding regression achieved a mean

squared error of 9.72 using the 210 cases without key missing values. So on the NPTS problem, RPR gives slightly better fit and is far more applicable.

5. CONCLUSIONS

This paper outlines the first steps in a process to determine the feasibility of using data fusion for understanding the full continuum of travel. We have focused in this portion of the research on the selection of matching variables, and have seen that RPR preformed well for describing local data variables and slightly less well in describing long-distance. We are not sure whether this is an anomaly of the datasets, and will give further attention.

Overall, this research provides a novel application of recursive partitioning methodology to travel survey data. The results are interpretable, and provide competitive fit when compared to conventional regression techniques. This approach enables automatic identification of key explanatory variables and interactions, which is valuable in terms of understanding the relation of variables in the data set. The partitions for each data set can be overlaid to produce a set of variables for matching in data fusion problems. This process will be the focus of further research.

6. ABOUT THE DATA

The Nationwide Personal Transportation Survey (NPTS) is a survey of typical daily travel performed by people in households all over the United States. All trips made during a pre-assigned 24-hour period by each household member five years of age and over in the sampled household were included in the survey. Details about the purpose of every trip, means of transportation, trip time and duration, number of household members and total number of people on the trip, driver, and vehicle characteristics are included in the data set.

The American Travel Survey (ATS) is a survey of long-distance travel made by people in households all over the United States. The ATS contains information about trips of 100 miles or more away from home taken by all modes of transportation for a period of one year. Details on origin, destination, purpose, and mode are included and the data provide insight into America's long-distance transportation choices, including foreign and domestic travel. Table 3 compares and contrasts the two surveys.

To ensure comparability between these surveys, we restrict attention to those households interviewed between July and December of 1995, inclusive, which is the overlapping period for reports. Also, we restricted our analyses to subjects who are 18 or older, because younger children are most often passengers, and thus more likely to contribute noise than signal.

Under these restrictions, we have NPTS data from 33,163 households and ATS data from 50,767 households. Both surveys recorded household demographic data (e.g., number and ages of residents, education levels, income, location, and so forth) as well as travel data for all modes of transportation.

Table 3 - Comparison of NPTS and ATS

	NPTS	ATS
Scope	All trips made by civilian, non-institutionalized persons living in households (not group quarters) who are aged five and older. Student dormitories were not included in the sample.	Trips to destinations 100 miles or more away from home made by civilian, non-institutionalized persons of all ages. College dormitories and housing were included in the sample of households.
Sample and population	List-assisted RDD/all U.S. households	Retired Census current population survey (CPS) address samples/all U.S. households
Mode of data collection	Computer assisted telephone interviewing (CATI),	Computer assisted telephone interviewing (CATI) and computer assisted person interviewing (CAPI).
Data collection period	One year: May 1995 through June 1996.	One year: April 1995 to March 1996.
Reporting period	1 day	3 months
Imputation	Data for missing values is coded as missing.	Values for missing data are estimated through imputation procedures. Imputed data are flagged.

For more information, please visit our websites at:

http://www-cta.ornl.gov/npts

http://www.bts.gov/programs/ats

ACKNOWLEDGEMENTS

The authors wish to thank Bernetta Crutcher, Mathematical Statistician at the Bureau of Transportation Statistics, who assisted in the CART analysis. She learned about and used two complex data sets, and ran the SAS analysis needed to make them meaningful. Without her effort and attention to detail this analysis could not have been completed.

REFERENCES

Baker, K., Harris, P. and O'Brien, J. (1994) Data Fusion: An Appraisal and Experimental Evaluation, **Journal of the Market Research Society**, Vol. 31, No. 2.

Breiman, L., Friedman, J., Olshen, R., and Stone, C., Classification and Regression Trees, 1984, The Wadsworth Statistics/Probability Series

Kamakura, W. and Wedel, M. (1996) Statistical Data-Fusion for Cross-Tabulation, Unpublished document, University of Pittsburgh.

Kish, L. (1999) Cumulating/Combining Population Surveys, **Survey Methodology**, Vol. 25, No. 2.

Madre, J.L and Maffre, J. (1999) Is it Necessary to Collect Data on Daily Mobility and on Long Distance Travel in the Same Survey?, **Proceedings Personal Travel: The Long and Short of It**, Washington, DC, June 1999.

Morgan, J.N., and Sonquist, J.A. 1963. Problems in the Analysis of Survey Data, and a Proposal. J. Amer. Statist. Assoc., 58: 415-434.

O'Brien, S. (1991) The Role of Data Fusion in Actionable Media Targeting in the 1990's, **Marketing and Research Today**, February 1991.

Radner, D., Allen, R., Gonzalez, M.E., Jabine, T.B., and Muller, H.J (1980) Report on Exact and Statistical Matching Techniques, **Statistical Policy Working Paper 5**, U.S. Department of Commerce, Washington, DC.: U.S. Government Printing Office.

Rassler, S. and Feischer, K (1999) Aspects Concerning Data Fusion Techniques, Unpublished document, Nurnberg.

Rodgers, W. (1984). An Evaluation of Statistical Matching **Journal of Business & Economic Statistics**, Vol. 2, No. 1, January 1984.

Rubin, D. (1976) Inference and Missing Data**, Biometrika**, 63, 581-592.

Rubin, D. (1986) Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations, **Journal of Business and Economic Statistics**, Vol. 4, No. 1.

Rubin, D. and N. Schenker (1986) Multiple Imputation for Interval Estimation From Simple Random Samples with Ignorable Nonresponse, **Journal of the American Statistical Association**, Vol. 81, No. 394, June 1986.

Shannon, W. and Banks, D. (1999) Combining Classification Trees Using MLE, **Statistics in Medicine**, Vol. 18, page 727-740.

Wiednebeck, M. (1999) Fusion of Data and Estimation by Entropy Maximization, Proceedings from Statistics Canada Symposium 99, Ottawa, Canada.