Studying patterns of use of transport modes through data mining - Application to U.S. national household travel survey data set

Studying Patterns of Use of Transport Modes through Data Mining--

Application to U.S. National Household Travel Survey Data Set

Marco Diana

28th February 2012

Preferred citation: this document may be cited directly referring to the above mentioned final published version:

Diana, M. (2012) Studying Patterns of Use of Transport Modes through Data Mining--Application to U.S. National Household Travel Survey Data Set, *Transportation Research Record: Journal of the Transportation Research Board, No. 2308*, pp. 1-9.

STUDYING PATTERNS OF USE OF TRANSPORT MODES THROUGH DATA MINING--APPLICATION TO U.S. NATIONAL HOUSEHOLD TRAVEL SURVEY DATA SET

Marco Diana*

Dipartimento di Ingegneria dell'Ambiente, del Territorio e delle Infrastrutture

Politecnico di Torino

Corso Duca degli Abruzzi, 24

10129 Torino – ITALY

Phone: +39 011 090 5638 - Fax: +39 011 090 5699

marco.diana@polito.it

Manuscript submitted for the 91st TRB annual meeting

Submission date: July 25, 2011 – Revised November 8, 2011 and February 29, 2012

Word count, excluding tables and figures: 6256

Number of tables and figures: 4 + 2

* *Corresponding author*

ABSTRACT

Travel-related data collection activities require high amounts of financial and human resources to be successfully carried out. In a context where the available resources are scarce, there is a need to exploit the information that is hidden in these dataset, to increase their added value and gain support among decision makers not to discontinue such efforts. The present research assesses the use of a data mining technique, Association Analysis, to better understand the patterns of mode uses from the 2009 U.S. National Household Travel Survey. Only variables related to self-reported levels of use of the different transportation means are considered, along with those useful to the socioeconomic characterization of the respondents. It has been possible to mine association rules that potentially show in economic terms a substitution effect between cars and public transportation, whereas such effect was not observed between public transportation and non-motorized modes (bike, feet). This is a policy relevant finding, since transit marketing should be targeted to car drivers rather than to bikers or walkers to really improve the environmental performances of any transportation system. Modal diversion from car to transit is seldom observed in practice, given the competitive advantage of private modes that has been extensively discussed in the literature. However, if we control for such factor, then our results suggest that modal diversion should mainly occur from cars to transit, rather than from non-motorized modes to transit.

KEYWORDS

Association rules, frequent itemsets, data mining, mode levels of use

INTRODUCTION

Running national travel surveys represents a true challenge, both in terms of mobilization of financial and human resources and of proper planning and organization of the related activities. In a context where funding opportunities are limited, it becomes therefore critical to be able to show the real added value of such data collection effort to the relevant decision makers, in order to gain their necessary support. This implies the necessity to better exploit these data, going beyond their traditional use for modeling purposes or for the compilation of descriptive statistics. On the other hand, analytical tools that are being used to study transportation problems nowadays rely on a wide range of techniques from different disciplines, from civil engineering to economics to social sciences. This "new generation" of analyses therefore continuously calls for innovating ways to use datasets, and this might represent a non-negligible stimulus for the implementation of future travel surveys, along with the need to keep on with the above mentioned more traditional exploitations of transportation data.

Within such context, the goal of the present paper is to assess Association Analysis as a tool that has been relatively little used since now to perform transportation research, in order to understand its usefulness to study complex mobility issues. Association Analysis is an emerging technique within the rapidly developing data mining research field (*1*,*2*), whose aim in broad terms is to find frequent combinations of values of non-metric variables in a dataset. Association Analysis is also named Association Rules Mining, Frequent Patterns Mining or Frequent Itemsets Mining, according to the specific methodology being adopted. Its more immediate application field is in marketing studies, where transactional databases, i.e. datasets where binary variables represent the purchases of goods, are mined to understand which items are more frequently bought together. However, this technique has been rapidly adopted in different fields, where large datasets are available and difficult to treat with multivariate statistical analyses. Successful applications include geographic information systems analysis, bioinformatics, text and multimedia documents mining or web structure analysis. Within transportation research, Association Analysis has been sporadically used in the last decade to study mobility behaviors with an activity-based approach (*3*,*4*), but also to forecast traffic flows (*5*) and to discover patterns of road accidents (*6*,*7*).

In the following we use this method for a different application, namely to study patterns of uses of different transportation modes, as they were reported by the respondents of the U.S. National Household Travel Survey (NHTS) that was completed in 2009. The importance of shedding more light on this issue is apparent for both researchers and practitioners. Modal diversion and co-modality are in fact consistently indicated as one of the most important ways to reduce congestion and environmental externalities of transportation, but related policies have been rather ineffective up to now. Like in the above mentioned purchase marketing studies, understanding if there are some modes that are more consistently used (or not used) together can be useful to promote the use of some modes and discourage the use of others through more targeted policies. In this sense, the present study can be seen as a complement to related research efforts in the transportation sector focusing for example on customer profiling and segmentation studies (*8*-*10*) or on subjective versus objective determinants of mode choice (*11*-*14*).

METHOD

We briefly recall the basic ideas that lie behind Association Analysis and that are more relevant to understand our subsequent application of the method, by giving an intuitive description of this method. We refer instead the interested reader to any textbook on data mining (e.g. *15*, *16*) for a more formal and complete presentation of the problem, the related solution techniques and algorithms, and the state of the art in this very active research field.

Association Analysis was originally proposed to understand which items are more often sold together in shops (*1*). As such, its input is a dataset in which each row is a transaction, i.e. a single purchase, and the columns are binary variables that indicate if the corresponding item has been bought during that transaction. However this scheme was soon adapted to study also different problems. In particular, rows could represent interviewed subjects and columns the corresponding responses or respondent's characteristics, as shown in Table 1. If responses are expressed through metric variables, boundary values can be used to redefine them as categorical ones, and these latter can in turn be transformed into binary variables through dummy coding, where each variable is set to 1 if the response of the subject falls within the corresponding category and 0 otherwise (for example, three income levels translate into the two dummies of columns 3 and 4 of Table 1). In this way, each subject is characterized by a specific combination of values of the binary dummy variables. An *itemset* is a set of items, i.e. a set of these binary variables with values equal to one for a specific subject or observation. Note that 2^n-1 itemsets can therefore be identified for each observation, if *n* items were "picked up" by the respondent (for example, considering again Table 1 there are three distinct itemsets for respondent 1). A *rule* is an implication linking together two itemsets within an observation. The first rule is called *antecedent*, or left-hand side (LHS) of the rule and the second one *subsequent*, or right-hand side (RHS). Following the definition in (*1*), we only consider rules with single items as subsequent. Therefore, one possible rule for respondent 2 of Table 1 could be written as follows: {FEMALE, MED_INCOME} => {TRANSIT_USE}.

Table 1 about here

Different measures are used to select the most relevant itemsets and rules. In particular, the *support* of an itemset (or of a rule) is the fraction of observations that contain the itemset (or the two itemsets of the rule), whereas the *confidence* of a rule is the fraction of observations containing the antecedent that also contain the subsequent. Another interestingness measure is the *lift*, defined as the ratio between the support of a rule and the product of the supports of the corresponding two itemsets. The greater the lift, the stronger the association between LHS and RHS is. Unlike correlations, significance levels or other goodness of fit measures in statistics, there is no theoretically grounded or practically recommended cutoff value that can be suggested for these measures, since they are used by the researcher simply to select the most interesting rules. Therefore, we present in the following different kinds of analyses, where rules are selected on the basis of widely different values of interestingness measures.

There are various reasons for which Association Analysis should be considered a useful addition to the set of tools of both transportation researchers and practitioners, particularly concerning the need of better exploiting the knowledge that is hidden in large transportation datasets. Recalling the framework that was described in the first paragraph of the Introduction, we note that this technique has been specifically conceived for a secondary use of data, i.e. to perform analyses and perhaps investigate issues that were initially not considered when

implementing the survey and gathering the data. This is a very effective way to increase the added value of a surveying activity and its support among stakeholders. Beyond this, off-the-shelf software packages are now available that make this tool accessible also to the larger research and practitioner community with a reasonable learning effort. In the following we will use *aRules*, an open source implementation of two popular Association Analysis algorithms (i.e. apriori and Eclat) that is available as a package of the *R* language (*17*).

Association Analysis, like other data mining techniques, is specially suited for handling very large datasets, with hundreds of variables and thousands of observations. Variables can be either metric or non-metric and no assumption is made concerning their distributions and relationships, since this is not a statistical analysis technique. Concerning this latter point, the other side of the coin is that such method is rather a-theoretical. Unlike most statistical techniques, it does not allow to make inferences or assess the degree of confidence of the patterns being found. Hence, this method is not well suited for building predictive models, and the assessment of the relevance of the patterns being mined must be made by the researcher on the basis of his/her knowledge of the topic.

Alternative methods from multivariate statistics are also available to study the interrelationships among non-metric and categorical variables. Cross-tabulations and scatterplots are the simplest one and easy to implement, but are of limited practical use when the analyst wants to jointly consider more than two-three variables. Multiple correspondence analysis is a powerful tool to discover patterns of values among categorical variables that has been previously used in transportation research (*18*, *19*), but also in this case the number of items must be limited in order to meaningfully interpret the results. Data mining techniques are capable of overcoming such limitations, thus allowing to analyze in an integrated way the large datasets that are available from a typical travel survey.

DATA

In the present research we focus our attention on a limited subset of variables from the 2009 NHTS person dataset. Such variables describe the levels of use of different modes, along with some key socio-economic indicators to characterize the respondents. We present these variables in Table 2, where the first two columns show their name and description, as they are reported in the survey documentation. The upper four variables in the table measure the levels of use of cars (as driver), public transportation, bikes and feet. Since these are metric variables, we had to individuate four to six categories for each one, in order to define a corresponding number of binary variables to be used in the analysis, according to the procedure reported in the previous section. These categories are reported in the last column of the table. Boundary values were set in order to broadly match the meaning of the corresponding category label, after having inspected the frequency distribution of the responses in order to avoid splitting peaks of such distribution in two or more categories.

In the following we would like to relate mode use patterns with the socioeconomic condition of the respondents. Past research efforts, such as the above mentioned customer profiling studies, have in fact shown the deep influence of such factors. Given the exploratory nature of the present study, that primarily aims at assessing the potentialities of a data mining method, we only consider two additional variables related to the socioeconomic status, which are reported in the last two rows of Table 2. Educational levels can be considered representative of the social class of the respondent, whereas travel-related medical conditions are a different factor

influencing mode uses patterns that is not directly linked to the former one. The categories of EDUC were grouped to have a sufficient number of observations in each group. Of course, socioeconomic determinants of mode choice form a long list beyond the two factors here considered, not to say of the role played by the performances of the different means. However, we focus only on two aspects that seem more difficult to consider in a traditional mode choice analysis, whereas the effect of other factors such as income and gender has been extensively studied.

Table 2 about here

Beyond the definition of categories from the original dataset, another preprocessing step was the exclusion of those respondents that did not answer to any of the four variables YEARMILE, PTUSED, NWALKTRP or NBIKETRP, including appropriate skips such as, for example, those aged less than 16 concerning YEARMILE. We also excluded all those declaring that public transportation is not available where they live. The rationale is in fact to study modal patterns of those that are in the condition of choosing between different transportation means, at least for some of their trips. The number of observations that we retained was therefore reduced to 123,020, from the 324,184 originally available, representing over 111 millions of U.S. residents out of 300 millions when considering observations weights. Descriptive statistics concerning these variables are available on the official NHTS website (http://nhts.ornl.gov/).

In order to assess the potentialities of the method, we remark that we did not consider trip-specific information and that we purposely selected those variables describing mode uses in more general terms. These latter are somewhat less informative when carrying out statistical analyses or a modeling exercise, since they are referred to general behaviors rather than to specific choices in a given situation. For the same reason, we consider here educational levels rather than other more “analyst-friendly” socioeconomic variables such as income. Variables from Table 2 are therefore often employed only to compile univariate descriptive statistics, that are surely interesting and informative but do not allow to uncover the underlying mode use patterns. A better exploitation also of this kind of information would increase the value of transport surveys datasets, and make it possible to more easily extract relevant information also from less specialized but still relevant mobility data sources, such as censuses, consumer expenditure surveys or time use surveys.

RESULTS

Rules with greatest support

After considerable preliminary analyses, the dataset described in the previous section was mined to find the rules with support greater than .01 and lift greater than 1.1., that are reported in Table 3. The support threshold is lower than the values usually being considered in the literature, due to the fact that some travel behaviors were seldom reported, particularly concerning the use of public transit or bikes. Therefore, considering higher support values would have implied to miss all those not exclusively using cars. On the other hand, we recall that minimum support is not related to the statistical concept of degree of significance with the related risk of considering spurious relationships, but it is only a cutoff value to focus the attention on a manageable number of rules that are more frequent in the dataset. This is an important point to stress, since the

software that we used to mine rules does not allow to treat unequal probability samples; therefore, support values that are shown in the table do not take into account observation weights. In any case, we also computed the "true" support of each rule by correcting for the weights after that the rules were mined. Departures from the values in the table range from -30% to +19%, an appreciable but not critical difference, given the role of support values in our framework.

Table 3 about here

Coming to the results, we preliminarily notice that all the rules that have been mined with the above criteria have PTUSED=None as RHS, and confidence values are always greater than .93. Therefore, this analysis could give a contribution in understanding which uses of alternative modes, levels of education and health conditions are more associated with the fact of not using transit. This is a very relevant research question, given the efforts that are conducted in many countries to promote the use of collective transportation means.

Rules in Table 3 are arranged in five different groups on the basis of the itemsets that they contain. In analogy with cluster analysis, we give an informative name to each group, in order to ease the interpretation of the results. The abbreviations of these names are shown in the last column of the table. Rules 1 to 4 show that people with a mean educational level that neither walk nor use bikes tend also not to use public transportation, even if they do not have a medical condition that makes it hard to travel. This result can be interpreted in light of the ongoing debate on the possibility of decreasing the modal market share of cars. In particular, funding transit improvements is generally more justifiable if car trips can be diverted, less so if bike or walk trips are instead substituted or new travel demand is induced. Putting it in economic terms, the existence of a substitution effect between car and other modes is desirable, whereas it is not desirable between transit and non-motorized transportation modes. Concerning this latter point, these four rules give a partially positive response, since they show a non-use of both transit and non-motorized modes for many respondents. However, it would be equally important to observe their joint utilization, in order to fully prove the absence of a substitution effect. This is not the case in the set of rules displayed in Table 3, mainly because the use of such modes is much less reported, so that the rules containing the corresponding items have a very low support. A different kind of analysis will therefore be reported in the following subsection to more thoroughly investigate on this point. Given the lack of use of bikes, walk and transit shown by this group of rules, they can be labeled as "Environmentally Insensitive" (EI).

The following six rules (5 to 10) can be interpreted along these same lines. We notice in fact that they associate the use of cars and the absence of use of non motorized modes with the fact of not using public transportation. Again, this is a necessary but not a sufficient condition to show substitution between cars and transit, so that more analyses are needed. Since we see here the use of cars, those rules are pointing to a larger environmental impacts of related mobility behaviors and we name them "Environmentally Harmful" (EH).

The group in the middle of the table (rules 11 to 15) replicates the results of the previous two groups, but for less educated people. Therefore, education levels seem not to affect mode use patterns. This seems a rather surprising result, given the well known effects of socioeconomic status on mobility levels. However we recall that here we rather focus on the combination of use of different transportation means rather than on their levels of consumption. In this case, contrasting trends could contribute in blurring differences among social groups. For example, the tendency of the poorer to use less private cars and more transit could be counterbalanced by the

fact of living in neighborhoods or territories that are less accessible by public transportation. We name such group "Less educated Replicas" (LR) of the above EI and EH rules.

The two rules of the following group are perhaps less interesting, since they simply characterize the group of less mobile (LM) people. The last group instead can shed some light on the use of different transportation means by those that declared having medical conditions that make it difficult to travel (they are about 12% of those that answered the related question). These mobility impaired (MI) people seem to consistently rely on cars, and their car use is not less frequent than that of the others.

To sum up, we observed through this set of rules some patterns that seem to point at the presence of substitution effects between cars and transit (i.e. the possibility that an increase in the use of one means induce the decrease in the use of the other), and at the absence of such effect between car and non-motorized modes. Yet we need to find rules that contain items related to the use of bikes and of public transportation in their RHS, in order to have more conclusive results. These rules cannot be mined simply considering a minimum support threshold, since the use of such modes is much less widespread in the sample.

Constraint-based association rules

In view of the previous initial results, we performed further investigations by mining rules that must contain an item related to the use of public transportation, bike or feet in their RHS. Support values are of course dramatically lower since we consider the use of less popular modes. This constraint-based association analysis is useful to shed light on such patterns of modal consumption that are less common in our sample, but of great interest for transport policy purposes.

Table 4 shows the 22 constrained rules that we obtained with support greater than 10^{-4}, but always lower than 10^{-3}. According to what was previously noticed, support values are much lower compared to the rules in Table 3 because modes other than cars are much less frequently picked up. This is a good example of the fact that there does not exist any pre-defined cutoff value for support, unlike most statistical analyses. On the other hand, all these rules have a confidence greater than 0.60 and a lift greater than 1.5. We preliminarily note that they all contain items related to a moderate to intensive use of bikes, transit and feet both in the LHS and in the RHS. This is another hint that points at the absence of a substitution effect among those environmentally benign modes.

Table 4 about here

As previously done, we can have a closer look at those rules by dividing them in five groups. The first one represents healthy subjects that walk a lot and either daily ride a bike or intensively use public transportation, thus showing a pro-environmental (PE) behavior. The following eight rules (from 33 to 40) contain instead items related to the use of cars in their LHS. Such items range from YEARMILE=Rare to YEARMILE=A_lot, and seem therefore to indicate that driving a car and walking are rather independent, at least when other modes are also used. We could therefore label these subjects as "multimodalists" (MU in the table), in the sense that they use all the available transportation modes.

The last three groups show different mode use patterns. Rule 41 is the only one where no item related to education appears, and represents those people that do not drive but daily walk ("Great Walkers" – GW). Rules 42-43 represent highly mobile (HM) and well educated

population segments, for which walking is a common activity but not a daily practice. Conversely, rules 44 and 45 trace back the characteristics of heavy public transportation users, that do not drive cars and walk a lot ("Transit Captives" – TC). The structure of this last group is complementary of that of group EH, where driving was associated to not using transit, and therefore reinforces the previous finding of a possible substitution effect between those two modes. Furthermore, the very high lift values for rules in groups HM and TC suggest that such clusters of individuals are quite well defined and distinct from the rest of the population.

Stability and transferability of the rules

The above analysis has been performed on a representative sample of the U.S. population and should have therefore a general validity. However, data mining techniques do not make any particular statistical assumptions. Therefore, both the stability of the results, for example in terms of their dependence on the definition of the categories that we adopted, and their transferability to a different population could be questioned. Concerning the former point, we repeated the above analyses by aggregating the less frequent categories that are shown in Table 2. The results were not significantly affected, beyond the fact of having to interpret a larger number of rules given the higher support of the corresponding itemsets. In particular, more constraint-based association rules showing some levels of use of transit in their RHS were mined. These latter rules generally confirmed and completed the ones reported in Table 2. In particular, it was more clearly shown that the fact of using transit is not associated to any particular level of use of bikes, thus reinforcing the above hypothesis of absence of substitution effects between bike and public transit modes.

Another analysis was performed to understand if the results would change when considering a different sample. For this, we mined the dataset from the previous 2001 NHTS, which contains the same variables that are reported in Table 2, although with some minor differences in their definition. Considering a sufficiently large set of rules with highest support allowed us to retrieve all those reported in Table 3. However, their respective support values become more dispersed, ranging from 0.01675 to 0.1323, with 55 rules in total having a support greater than this lower value. This is essentially due to the different relative frequencies concerning educational levels and medical conditions. In particular, the 2009 sample presents a lower proportion of less educated persons (up to high school) and a higher proportion of people that declared having a troubling medical condition, which is probably largely due to population aging. To sum up, when controlling for those factors, that are linked to well-known demographic trends in most Western countries over the last decade, the above results could still be observed.

Constraint-based association analysis is much more sensitive to changes in the distribution of the support values in the sample, since it involves mining rules with rare items, whose frequency can dramatically change for even small variations in the composition of the sample in absolute terms. Therefore, results displayed in Table 4 could only partially be replicated by using the 2001 sample. In particular, the four constraint-based rules with highest support associated a moderate use of bikes with an heavy use of feet, like rules of group F, but without containing items related to the use of public transportation. In other words, although the itemsets of rules were not identical, their meaning and interpretation is much the same in both datasets.

GEOGRAPHIC CHARACTERIZATION OF THE GROUPS

The groups of rules that were found in the preceding section can be associated with a corresponding subset of respondents in the sample. We can therefore study if people residing in different areas of the country have different patterns of use of transportation modes. For this, we consider the variable CDIVMSAR from the dataset, which categorizes households on the basis of the nine census divisions, the fact of being or not located in a metropolitan statistical area (MSA), above or below 1 million inhabitants and in presence or absence of a subway system. The characterization of our groups by census division is reported in Figure 1 and the one by MSA status and presence of subway in Figure 2. For comparative purposes, the two charts also show the same information for both the whole U.S. population, according to the entire NHTS sample, and for the selected sample of individuals that were considered in this study (as we noted, these latter are about one third of the total). Both figures keep into consideration observation weights.

Figures 1 and 2 about here

Groups EI and EH have the same geographical characterization of the U.S. population considering both census divisions and MSA status, thus reflecting modal consumption behaviors that are uniformly distributed across the U.S., whereas the characterization of group LR reflects differences in educational levels of residents in different divisions (*20*). Our selected sample instead is biased towards larger MSAs with subways, an expected result recalling that we did not consider those respondents that declared not having public transportation services available where they live. On the other hand, group LM shows a higher proportion of less mobile people in MSA with less than 1 million inhabitants. Group MI captured people that have difficulties in traveling and that are exclusive and regular car users. It is interesting here to note that their distribution in the territory is not substantially different from that of the entire population, despite the fact that competing transportation means, particularly public transportation, have a different appeal at least according to MSA status. This finding seems therefore to suggest that the reliance on cars for this population segment is not substantially affected by the quality of the offer of alternative means.

The groups shown in table 4 show instead rather diverging geographical compositions, reflecting the fact of representing some “market niches”. Group PE, i.e. people walking a lot and also using either public transportation or bikes, is of course more concentrated in larger metropolitan areas and in the corresponding census divisions. A similar pattern can be detected for group MU, i.e. those that also use cars beyond the above means, although we find there a larger proportion of residents in the South Atlantic division and in larger MSA without subway. We labeled such group as “multimodalists”, and therefore it should be relatively easier for them to reduce the use of cars in presence of a competing offer of those modes that they already use. Therefore, policy actions aimed at promoting public transportation and bikes and targeted to such group could be quite effective in such geographical regions.

Groups GW and HM show a widely different distribution compared to that of the other groups. However we see from Table 4 that they are defined from a small number of rules with low support; in particular, only 17 and 13 unweighted observations respectively belong to those groups. These figures therefore are not enough stable and too strongly dependent on single observations and on their corresponding weights. The cardinality of the last group TC is higher (52 observations representing more than 210,000 individuals) and we notice there a high number of persons residing in the Mid Atlantic division and in a larger MSA with subway. This figure

shows a very high concentration of transit dependents in the New York metropolitan area, also compared to other major cities in the U.S., where transit and cars are more often both used.

CONCLUSIONS

In this paper we presented some analyses aimed at uncovering mode use patterns within the sample of the U.S. National Household Travel Survey administered in 2009. We focused on a small subset of variables that report the levels of use of different modes, and we showed that it is possible to extract interesting information even from such more general variables, that are generally disregarded in more complex analyses. Association Analysis has proven itself a useful tool to achieve this goal, given its ability to treat large datasets without making any statistical assumption on the variables being analyzed. This flexibility allows the researcher to consider factors affecting travel behavior that are seldom investigated since they are hard to quantify, such as health conditions. On the other hand, Association Analysis is an exploratory technique where the analyst has little guidance on how to select the most interesting rules. Measures such as support can give some hints but cannot be the only criterion, especially when some items are much less frequently picked up by respondents and nevertheless related patterns are of interest, such as transit versus car use in the NHTS dataset.

The set of rules that has been mined gives some initial support to the existence of a substitution effect between cars and transit, and to the absence of such substitution effect between transit and non-motorized modes, since Association Analysis has found some necessary albeit not sufficient conditions for such effect. If confirmed, this would be an important result for policy makers aiming to achieve modal diversion, since policies aimed at expanding transit offer should be able to substitute more car trips than other trips done with non motorized modes. However, the existence of such mechanism does not guarantee that modal diversion can truly be reached, since a lot of other factors, that are both instrumental and affective and that were not considered here, come into play and make practically quite hard to substitute car trips with transit trips (*14*). Our main conclusion is that, if such factors can be properly addressed, then we would probably see more modal switches to transit from cars than from bikes and feet.

A rather straightforward extension of the present research would imply the use of Association Analysis to analyze trips recorded in the NHTS dataset. In particular, a promising research avenue consists in using a sequential pattern mining algorithm, that is a generalization of the analysis technique that we used here to treat panel data. In this way, patterns of mode use could be linked to trip-specific rather than to person-related factors. Beyond this point, in this preliminary research we chose to geographically characterize groups after that they were defined through data mining technique. An alternative methodology would consist in directly using the corresponding categorical variables in the mining algorithm. This would probably give a less complete overview on the relationships among geographical factors and mode use patterns; on the other hand, it would probably highlight some specific and territory-related patterns that could not be detected here.

ACKNOWLEDGEMENTS

We acknowledge the helpful comments of five referees on a previous version of this paper.

REFERENCES

1. Agrawal, R. T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, D.C., 1993, pp. 207-216. Available at http://doi.acm.org/10.1145/170035.170072 – Accessed February 29th, 2012.
2. Han, J., H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, vol. 15(1), 2007, pp. 55-86.
3. Keuleers, B., G. Wets, T. Arentze, and H. Timmermans. Association rules in identification of spatial-temporal patterns in multiday activity diary data. *Transportation Research Record 1752*, 2001, pp. 32-37.
4. Keuleers, B., G. Wets, H. Timmermans, T. Arentze, and K. Vanhoof. Stationary and time-varying patterns in activity diary panel data: explorative analysis with association rules. *Transportation Research Record 1807*, 2002, pp. 9-15.
5. Gong, X., and X. Liu. A data mining based algorithm for traffic network flow forecasting. In *Proceedings of the IEEE International Conference on Integration of Knowledge Intensive Multi-Agent Systems* (KIMAS), Boston (MA), 2003, pp. 243-248. Available at http://dx.doi.org/10.1109/KIMAS.2003.1245052 – February 29th, 2012.
6. Geurts, K., I. Thomas, and G. Wets. Understanding spatial concentrations of road accidents using frequent item sets. *Accident Analysis & Prevention*, vol. 37(4), 2005, pp. 787-799.
7. Pandel, A., and M. Abdel-Aty. Discovering indirect associations in crash data through probe attributes. *Transportation Research Record 2083*, 2009, pp. 170-179.
8. Jensen, M. Passion and heart in transport—a sociological analysis on transport behavior. *Transport Policy*, vol. 6(1), 1999, pp. 19-33.
9. Anable, J. 'Complacent car addicts' or 'aspiring environmentalists'? Identifying travel behaviour segments using attitude theory. *Transport Policy*, vol. 12(1), 2005, pp. 65-78.
10. Diana, M., and P.L. Mokhtarian. Grouping travelers on the basis of their different car and transit levels of use. *Transportation*, vol. 36(4), 2009, pp. 455-467.
11. Vredin Johansson, M., T. Heldt, and P. Johansson. The effects of attitudes and personality traits on mode choice. *Transportation Research A*, vol. 40(6), 2006, pp. 507-525.
12. Scheiner, J., and C. Holz-Rau. Travel mode choice: affected by objective or subjective determinants?. *Transportation*, vol. 34(4), 2007, pp. 487-511.
13. Domarchi, C., A. Tudela, and A. Gonzales. Effect of attitudes, habit and affective appraisal on mode choice: an application to university workers. *Transportation*, vol. 35(3), 2008, pp. 585-599.
14. Diana, M. From mode choice to modal diversion: a new behavioural paradigm and an application to the study of the demand for innovative transport services. *Technological Forecasting and Social Change*, vol. 77(3), 2010, pp. 429-441.
15. Tan, P.-T., M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson Addison Wesley, Boston, 2006.
16. Nisbet, R., J. Elder, and G. Miner. *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press, Burlington, 2009.
17. Hahsler, M., B. Grün, K. Hornik, and C. Buchta. *Introduction to arules—A computational environment for mining association rules and frequent item sets*. Mimeo, 2010. Available at http://cran.r-project.org/web/packages/arules/vignettes/arules.pdf – Accessed February 29th, 2012.

18. Golob, T.F., and D.A. Hensher. The trip chaining activity of Sydney residents: A cross-section assessment by age group with a focus on seniors. *Journal of Transport Geography*, vol. 15(4), 2007, pp. 298-312.
19. Diana, M., and C. Pronello. Traveler segmentation strategy with nominal variables through correspondence analysis. *Transport Policy*, vol. 17(3), 2010, pp. 183-190.
20. U.S. Census Bureau. *2006-2008 American Community Survey 3-Year Estimates – Table S1501 (by Census Division)*. This table can be built online from http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml – Accessed February 29th, 2012.

List of figures

List of tables

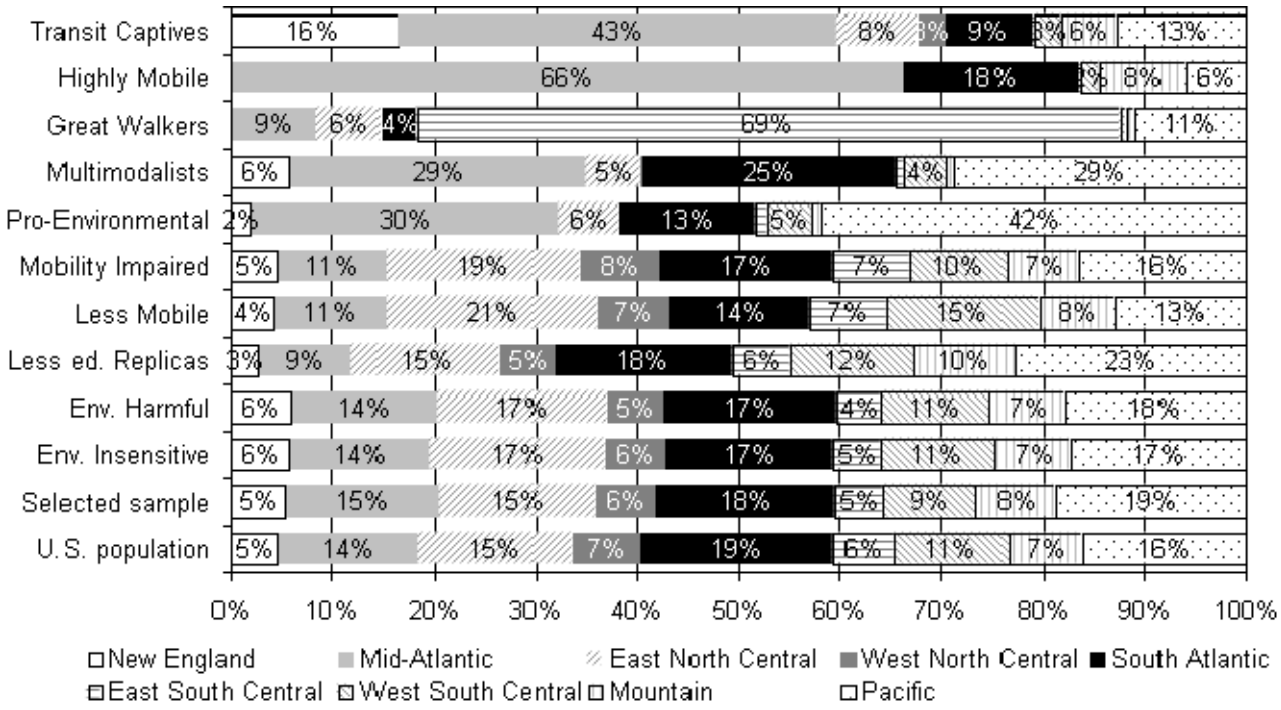

FIGURE 1 Census division by rules groups.

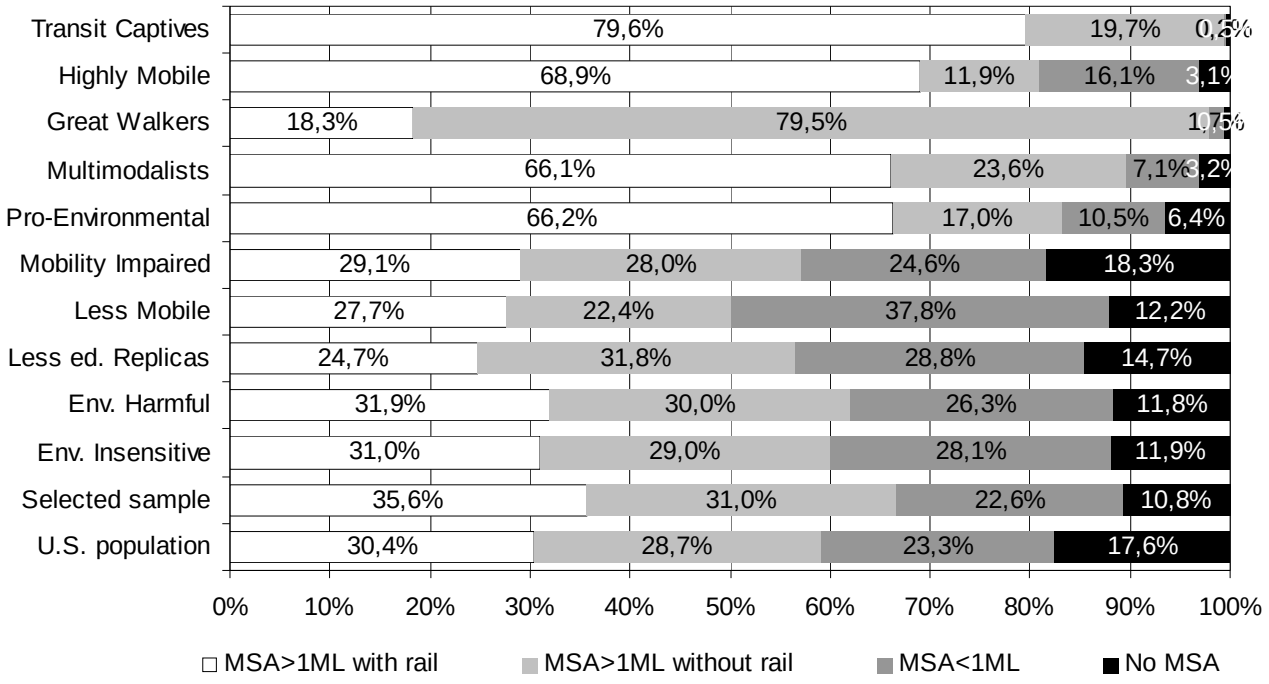

FIGURE 2 Metropolitan Statistical Area and presence of subway by rules groups.

TABLE 1 Hypothetical Example of an Association Analysis Database Built from a Mobility Survey

Respondent	Female	Med_ inc.	High_inc.	Health_cond.	Car_use	Transit_use
1	0	0	0	1	0	1
2	1	1	0	0	0	1
3	1	1	0	0	1	1
4	0	0	1	0	1	0

TABLE 2 List of the Considered Variables and Corresponding Categories

Variable	Description	Categories
YEARMILE	Miles driven during the past 12 months	None (0); Rare (1-50); Few (51-1000); Normal (1001-12000); A_lot (More than 12000)
PTUSED	Number of times transit was used in past month	None (0); Few (1-2); Weekly (3-10); A_lot (More than 10)
NBIKETRP	Number of bike trips in past week	None (0); Few (1-2); Daily (3-7); A_lot (More than 7)
NWALKTRP	Number of walk trips in past week	None (0); Few (1-2); Daily (3-7); A_lot (More than 7)
EDUC	Highest grade completed	No_diploma; High_school; College; Bachelor; Graduate
MEDCOND	Have medical condition making it hard to travel	Hard_to_travel; All_right

TABLE 3 Rules with Support > 0.01 and Lift > 1.1

N	LHS		RHS	Supp.	Conf.	Lift	Group
1.	{EDUC=High_school, NWALKTRP=None}	=>	{PTUSED=None}	0.085	0.945	1.108	EI
2.	{EDUC=High_school, NBIKETRP=None, NWALKTRP=None}	=>	{PTUSED=None}	0.083	0.946	1.109	EI
3.	{EDUC=High_school, MEDCOND=All_right, NWALKTRP=None}	=>	{PTUSED=None}	0.072	0.944	1.107	EI
4.	{EDUC=High_school, MEDCOND=All_right, NBIKETRP=None, NWALKTRP=None}	=>	{PTUSED=None}	0.070	0.945	1.108	EI
5.	{EDUC=High_school, NWALKTRP=None, YEARMILE=Normal}	=>	{PTUSED=None}	0.049	0.950	1.114	EH
6.	{EDUC=High_school, NBIKETRP=None, NWALKTRP=None, YEARMILE=Normal}	=>	{PTUSED=None}	0.048	0.951	1.115	EH
7.	{EDUC=High_school, MEDCOND=All_right, NWALKTRP=None, YEARMILE=Normal}	=>	{PTUSED=None}	0.042	0.949	1.113	EH
8.	{EDUC=High_school, NWALKTRP=None, YEARMILE=A_lot}	=>	{PTUSED=None}	0.021	0.942	1.105	EH
9.	{EDUC=High_school, NBIKETRP=None, NWALKTRP=None, YEARMILE=A_lot}	=>	{PTUSED=None}	0.021	0.942	1.104	EH
10.	{EDUC=High_school, MEDCOND=All_right, NWALKTRP=None, YEARMILE=A_lot}	=>	{PTUSED=None}	0.020	0.943	1.106	EH
11.	{EDUC=No_diploma, NBIKETRP=None, YEARMILE=Normal}	=>	{PTUSED=None}	0.021	0.939	1.101	LR
12.	{EDUC=No_diploma, NWALKTRP=None}	=>	{PTUSED=None}	0.017	0.949	1.113	LR
13.	{EDUC=No_diploma, NBIKETRP=None, NWALKTRP=None}	=>	{PTUSED=None}	0.017	0.949	1.113	LR
14.	{EDUC=No_diploma, MEDCOND=All_right, NWALKTRP=None}	=>	{PTUSED=None}	0.013	0.950	1.114	LR
15.	{EDUC=No_diploma, MEDCOND=All_right, NBIKETRP=None, NWALKTRP=None}	=>	{PTUSED=None}	0.013	0.950	1.114	LR
16.	{EDUC=High_school, NWALKTRP=None, YEARMILE=Few}	=>	{PTUSED=None}	0.011	0.941	1.104	LM
17.	{EDUC=High_school, NBIKETRP=None, NWALKTRP=None, YEARMILE=Few}	=>	{PTUSED=None}	0.011	0.944	1.107	LM
18.	{MEDCOND=Hard_to_travel, NWALKTRP=None, YEARMILE=Normal}	=>	{PTUSED=None}	0.021	0.946	1.109	MI
19.	{MEDCOND=Hard_to_travel, NBIKETRP=None, NWALKTRP=None, YEARMILE=Normal}	=>	{PTUSED=None}	0.021	0.947	1.111	MI
20.	{EDUC=High_school, MEDCOND=Hard_to_travel, NWALKTRP=None}	=>	{PTUSED=None}	0.013	0.950	1.114	MI
21.	{EDUC=High_school, MEDCOND=Hard_to_travel, NBIKETRP=None, NWALKTRP=None}	=>	{PTUSED=None}	0.013	0.951	1.115	MI
22.	{EDUC=High_school, MEDCOND=Hard_to_travel, YEARMILE=Normal}	=>	{PTUSED=None}	0.013	0.940	1.102	MI
23.	{EDUC=High_school, MEDCOND=Hard_to_travel, NBIKETRP=None, YEARMILE=Normal}	=>	{PTUSED=None}	0.013	0.941	1.104	MI

(Blank page)

TABLE 4 Rules with the use of feet or transit in their RHS

N	LHS		RHS	Supp.	Conf.	Lift	Group
24.	{EDUC=Bachelor, MEDCOND=All_right, NBIKETRP=Few, PTUSED=Weekly}	=>	{NWALKTRP=Daily}	7.2^{-4}	0.650	1.68	PE
25.	{EDUC=College, MEDCOND=All_right, NBIKETRP=Few, PTUSED=Weekly}	=>	{NWALKTRP=Daily}	3.7^{-4}	0.613	1.59	PE
26.	{EDUC=College, MEDCOND=All_right, NBIKETRP=Few, PTUSED=A_lot}	=>	{NWALKTRP=Daily}	2.0^{-4}	0.610	1.58	PE
27.	{EDUC=High_school, MEDCOND=All_right, NBIKETRP=Few, PTUSED=Few}	=>	{NWALKTRP=Daily}	2.5^{-4}	0.608	1.57	PE
28.	{EDUC=High_school, MEDCOND=All_right, NBIKETRP=Daily, PTUSED=Weekly}	=>	{NWALKTRP=Daily}	1.3^{-4}	0.615	1.59	PE
29.	{EDUC=High_school, MEDCOND=All_right, NBIKETRP=Few, PTUSED=A_lot}	=>	{NWALKTRP=Daily}	1.2^{-4}	0.714	1.85	PE
30.	{EDUC=Bachelor, NBIKETRP=Few, PTUSED=Weekly}	=>	{NWALKTRP=Daily}	7.2^{-4}	0.645	1.67	PE
31.	{EDUC=High_school, NBIKETRP=Daily, PTUSED=Weekly}	=>	{NWALKTRP=Daily}	1.5^{-4}	0.621	1.60	PE
32.	{EDUC=High_school, NBIKETRP=Few, PTUSED=A_lot}	=>	{NWALKTRP=Daily}	1.2^{-4}	0.682	1.76	PE
33.	{EDUC=Bachelor, NBIKETRP=Few, PTUSED=Few, YEARMILE=A_lot}	=>	{NWALKTRP=Daily}	4.6^{-4}	0.602	1.56	MU
34.	{EDUC=Bachelor, NBIKETRP=Few, PTUSED=Weekly, YEARMILE=Normal}	=>	{NWALKTRP=Daily}	3.7^{-4}	0.662	1.71	MU
35.	{EDUC=Bachelor, NBIKETRP=Few, PTUSED=Weekly, YEARMILE=A_lot}	=>	{NWALKTRP=Daily}	3.3^{-4}	0.651	1.68	MU
36.	{EDUC=Bachelor, NBIKETRP=Few, PTUSED=None, YEARMILE=Few}	=>	{NWALKTRP=Daily}	1.1^{-4}	0.667	1.72	MU
37.	{EDUC=College, NBIKETRP=Few, PTUSED=Few, YEARMILE=Normal}	=>	{NWALKTRP=Daily}	2.8^{-4}	0.618	1.60	MU
38.	{EDUC=College, NBIKETRP=Few, PTUSED=Weekly, YEARMILE=Normal}	=>	{NWALKTRP=Daily}	2.4^{-4}	0.714	1.85	MU
39.	{EDUC=College, NBIKETRP=None, PTUSED=A_lot, YEARMILE=Rare}	=>	{NWALKTRP=Daily}	1.7^{-4}	0.600	1.55	MU
40.	{EDUC=High_school, NBIKETRP=Few, PTUSED=Few, YEARMILE=Normal}	=>	{NWALKTRP=Daily}	1.5^{-4}	0.613	1.58	MU
41.	{MEDCOND=All_right, NBIKETRP=Few, YEARMILE=None}	=>	{NWALKTRP=Daily}	1.4^{-4}	0.630	1.63	GW
42.	{EDUC=Graduate, MEDCOND=All_right, NBIKETRP=A_lot, YEARMILE=A_lot}	=>	{NWALKTRP=A_lot}	1.1^{-4}	0.619	4.95	HM
43.	{EDUC=Graduate, NBIKETRP=A_lot, YEARMILE=A_lot}	=>	{NWALKTRP=A_lot}	1.1^{-4}	0.619	4.95	HM
44.	{EDUC=Graduate, MEDCOND=All_right, NWALKTRP=A_lot, YEARMILE=None}	=>	{PTUSED=A_lot}	1.5^{-4}	0.720	20.94	TC
45.	{EDUC=College, MEDCOND=All_right, NWALKTRP=A_lot, YEARMILE=None}	=>	{PTUSED=A_lot}	2.8^{-4}	0.654	19.02	TC

(blank page)