A study of tour-based mode choice based on a Support Vector Machine classifier

A STUDY ON TOUR-BASED MODE CHOICE BASED ON A SUPPORT VECTOR MACHINE CLASSIFIER

Miriam Pirra, Marco Diana

1st November 2018

A study on tour-based mode choice based on a Support Vector Machine classifier

Miriam Pirra[a] and Marco Diana[a]

[a] *DIATI - Department of Environment, Land and Infrastructure Engineering, Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino, ITALY*

Corresponding author: Miriam Pirra. Email: miriam.pirra@polito.it

A study on tour-based mode choice based on a Support Vector Machine classifier

A new approach in recognizing travel mode choice patterns based on a classification technique named Support Vector Machine is proposed. The tour-based travel demand dataset analysed is derived from the 2009 National Household Travel Survey. The main features characterizing each tour are the means used, travel-related variables and socioeconomic aspects.
Results obtained demonstrate the ability in predicting to some extent, in a real settings where car use dominates, which tours are likely to be made by public transport or non-motorized means. Moreover, the flexibility of the technique allows assessing the predictive power of each feature according to the combination of travel means used in different tours. Potential applications range from activity-based travel choices simulators to search engines supporting personalized travel planners, in general whenever "best guesses" on mode choice patterns have to be quickly made on large amount of data prejudicing the possibility of setting up a statistical model.

Keywords: trip chain; tour; mode choice; Support Vector Machine; multimodality; New York State

Introduction

Mode choice is certainly one of the main important aspects in travel behaviour analysis and the many variables that can influence the choice are key elements to be analysed. Individual and personal attributes, such as socioeconomic characteristics, are surely fundamental and they have to be combined with more general aspects related to the travel itself, as its purpose or duration.

Various solutions to the mode choice problem have been developed, many of which proposing more and more advanced formulations of disaggregated discrete choice models whose foundations stretch back 40 years by now (McFadden 1973; Daly and Zachary 1978). While such econometric formulations based on the random utility theory represent the state of the art in this field, alternative research approaches started

being considered in more recent years. In particular, the development of the data mining and machine learning domain has opened new research avenues since the late 90s also in the transport field. Mode choice modelling can thus be seen as a pattern recognition problem, with a more detailed analysis on the changing influence of each exogenous variable on different categorical outcomes, which is something different from the estimation of a unique coefficient and significance level for each variable.

A growing body of literature compares the performances of econometric models with different variants of data mining and machine learning methods (Xie, Lu, and Parkany 2003; Beelen, Thomas, and Verhetsel 2005; Omrani et al. 2013; Seyedabrishami and Shafahi 2013; Xian-Yu 2011). For example, the machine learning technique called Artificial Neural Network represents a valid option in many problem instances compared to Multinomial Logit models (Nijkamp, Reggiani, and Tritapepe 1996; Hensher and Ton 2000; Shukla et al. 2013). Recent researches show an increasing interest in the application of data mining techniques to mode choice prediction. For example, Omrani presents a study on daily trips in the city of Luxembourg through the comparison of results obtained using four methods: Artificial Neural Net-MLP, Artificial Neural Net-RBF, Multinomial Logistic Regression and Support Vector Machines (SVM) (Omrani 2015). However, multimodal combinations are not considered.

While trip-based mode choice methods such as those reviewed so far are mostly used in practical settings, tour-based approaches have clear advantages in that they allow for the explicit consideration of constraints induced by trip chains. Two papers could be found that study tour based mode choice through data□driven methodologies based on fuzzy sets (Shukla et al. 2013; Shukla et al. 2015) while an interesting relationship between trip chaining and mode choice is presented in Islam and Habib

(2012).

Among such methods, SVM has been sporadically adopted in the transportation research domain and the few applications that are related to mode choice actually consider trip-level analysis. For example, SVM has been related to travel mode choice modelling in the case of data collected in the San Francisco Bay Area in California (Zhang and Xie 2008). Another interesting analysis is found in Xian-Yu (2011), which studies modal choice in work related trips. The work presented in Yang et al. (2010), instead, concentrates on trip chains, but the goal is the recognition of activity features. More widely, SVM appears in the transportation field as predictor of travel time (Vanajakshi and Rilett 2007; Wu, Ho, and Lee 2004), incident detection (Kim, Lee, and Cho 2007) or net flow forecasting (Cheu et al. 2006). In relation with such state of the art, this paper applies SVM to the study of tour-based mode choice, to the best of authors' knowledge for the first time in the open literature. The goal is to assess the added value of such method for the problem under consideration, particularly concerning the role that different personal and tour-related characteristics have in shaping the pattern recognition problem.

This article is structured as follows. Support Vector Machine, a technique widely used to study classification problems, is firstly introduced. Then, the description of the tours dataset for New York State considered to analyse the modal behaviour of people travelling during a day is provided. After the exposition and the discussion of the most interesting results, conclusions and possible future works are presented.

Support Vector Machine classifier

Support Vector Machine (SVM) is a well-known computational learning method developed in the 80s and widely used for data classification and regression (Vapnik 1982). The aim of this technique is the analysis of data and the recognition of certain

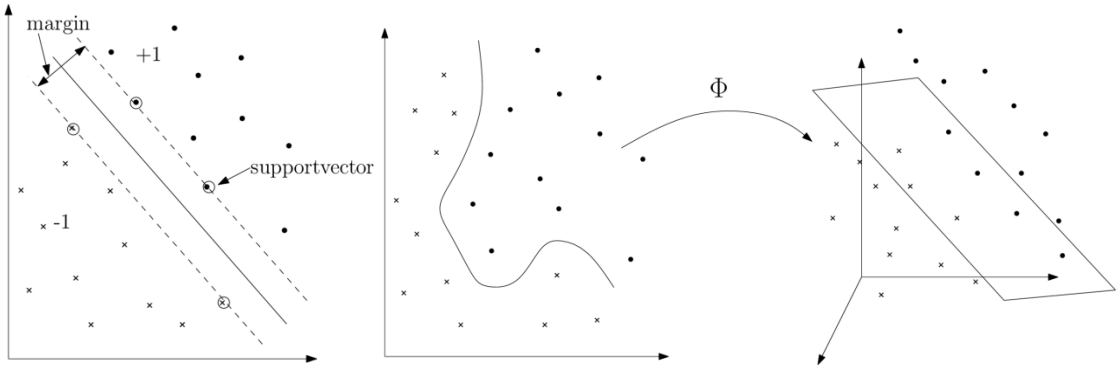
patterns that could be, then, used to classify unknown elements. Beyond its use in the transportation domain that was earlier reviewed, SVM has been widely and successfully adopted for solving classification problems in many different fields such image classification, text categorization, medical science or mechanical machine diagnostics.

The usual example proposed in machine learning to better explain how the various methods work refers to the Iris dataset (Tan, Steinbach, and Kumar 2005). This example is used in the following for a better comprehension of the technique described. The dataset includes 150 flowers belonging to the Iris kind, equally partitioned in three different species: Setosa, Versicolour and Virginica. Four botanical features characterizing these plants are known: sepal length, sepal width, petal length and petal width. In a classification procedure, the goal is to assign a new unknown Iris to one of the species, since they are properly described by those four features.

The simplest case is when the dataset is made up of only two classes, as could be, in the example, having only Setosa and Versicolour flowers. SVM technique customarily assigns two different labels to these classes: positive, "+1", and negative, "-1". The next step requires the definition of the *n*-dimensional "feature space", where *n* is number of attributes that are relevant to capture the pattern under investigation. In the Iris example, *n* is equal to four, which corresponds to the elements used to characterize the flowers (sepal length, sepal width, petal length and petal width). Thus, the dataset that the classifier has to analyse is made up of *M* elements, namely *M* vectors x_i, $i = 1, 2, \ldots, M$, where each vector has *n* dimensions.. In the example considered, if only two classes are considered, M is equal to 100 (50 Setosa and 50 Versicolour).

The SVM technique aims at creating an *n*-dimensional hyperplane separating the two groups and, therefore, acting as a classifier when some new points are given as test. Thus, a fraction of the *M* observations in the dataset (usually 75-80% of elements for

each class) is used to "train", i.e. build, the classifier. Therefore, each observation in the "train" subset belongs to a unique known class and it contributes to the creation of a proper boundary. To have the best classification, this boundary must be placed such that its distance from the nearest data points in each class is maximal; for this reason, its name is maximum margin (Figure 1(a)). These latter points are called support vectors, hence the name of this method. They are the core of the technique since they contain all the information necessary to properly define the classifier. . The elements not included in the "training" part are instead used as "test": a decision function assigns a label to these points according to their relative position from the separating hyperplane. Since they are originally assigned to a class (in fact they come from the remaining 25-20% of the dataset), it is possible to verify the ability of the SVM to properly identify their belonging.



(a) (b)

Figure 1. SVM representation of the two classes (labelled "+1" and "-1"), the maximum margin and the support vectors (a) and mapping of the non-linear case to a feature space where the boundary is linear (b).

One of the main advantages of this method stands in its application also to non-linear classification cases, that is, when the division boundary could not be defined through a linear relationship. In fact, SVM can use a more generic function $\phi(x_i)$ that

maps the data x_i to be classified onto a high-dimensional feature space, where the linear classification can then be adopted. Figure 1(b) shows the non-linear boundary, analytically not easy to define, that becomes, thanks to the projection to a higher-dimensional space, a linear hyperplane, computationally and mathematically more manageable. Moreover, thanks to a kernel function $K(x_i,x_j) = (\phi^T(x_i) \cdot \phi(x_j))$, the previous function $\phi(x_i)$ does not need to be evaluated, leading to a reduction of computational problems (Burges 1998). The kernel function used in this work is the Radial Basis Function (RBF):

$$K(x_i, x_j) = \exp(-\gamma \left\|x_i - x_j\right\|^2), \gamma > 0.$$

Support Vector Machine is, by definition, a two-class classifier, but it can be used for the solution of multi-classes problems too, as in the original Iris example. One of the possible approaches, which is also used in this paper, is the One-against-all (OAA): a separate SVM is created for each class and these elements are used as training data against all the others. Then, the assignation of the instances to the classes could be done in different ways. An intuitive approach sees outputs of the binary classifiers as votes and selects the class getting most votes as good. Further details can be found in Hsu and Lin (2002).

Dataset and analysis of tours complexity

The dataset analysed in this article is derived from the 2009 National Household Travel Survey (NHTS) public use files, which provide information on daily trips in the United States (U.S. Department of Transportation 2009). The focus is on the interviews gathered in New York State, where a survey add-on was financed and therefore a relatively high number of observations are available. In this way, a more balanced distribution of modes used in travelling is found, due to the strong use of public

transport in urbanized areas within that State. The first step involves the identification of all the tours that have been made by the survey respondents during the surveyed day. A tour is formally defined as a sequence of trips starting and ending at the same location (Axhausen 2008). Usually, the reference location is the house and, so, tours starting and ending at home within the surveying period are investigated. In the NHTS dataset this kind of sequences are found in the "Day Trip" file to produce the so called "Home-Based (HB) tours" list (Pirra and Diana 2016).

The analysis focuses on three main categories of travel means: individual motorized means (IM), public transport (PT), and bicycle and walking, which are jointly considered (BW). Thus, only tours made using these transport modes or their combinations are selected. Starting from the list of recorded trips, 39,167 tours made by 24,396 individuals on their survey day are finally reconstructed.

On the other hand, the analysis of the main purpose of each tour leads to the partition of the dataset in seven main categories with different sizes. Table 1 reports, in the first three rows, the unweighted number of tours (equalling to the number of observations in the dataset), the weighted percentage (computed considering observation weights for unbiased estimates) and the mean number of trips composing the tours for each group, irrespective of the mode used. Work and education are the unique purpose respectively of 10.40% and 4.89% of tours, which are then labelled with HWH and HEH. Among all the other possible purposes, a categorization is defined on personal activities (leading to HPH tours that only contain such activities), on social activities (HSH) or those whose goal is to transport someone (HTH). The first case collects 22.73% of all tours and includes shopping and personal care, while the second refers to all kind of meetings, meals, sport-related activities and friends and relatives visits (18.59% of tours). Activities such as transporting, picking up and dropping off

someone finally denote a smaller group of tours (3.99%). Tours starting and ending at home with no intermediate activity, i.e. going for a walk, jogging, taking the dog out and similar, are marginal (HH). However, the largest number of tours combines more than one of the above kinds of activities and they will be indicated as "HxH tours".

Table 1. Cross-tabulation of Home Based tours in New York State by travel modes and activities: occurrences, weighted percentages and mean number of trips per tour.

Means		HWH	HEH	HPH	HSH	HTH	HH	HxH	All activities
All modes	Nb. tours	4,672	1,909	10,404	9,829	1,995	149	10,209	39,167
	(row %)	(10.40)	(4.89)	(22.73)	(18.59)	(3.99)	(0.18)	(39.22)	(100.00)
	Av. trips	2.3	2.1	2.7	2.3	2.2	1.0	4.7	3.3
IM	Nb. tours	4,066	506	8,489	7,075	1,849	89	8,276	30,350
	(column %)	(65.19)	(25.30)	(60.14)	(52.71)	(84.29)	(42.87)	(59.25)	(58.16)
	Av. trips	2.3	2.0	2.8	2.2	2.2	1.0	4.6	3.3
PT	Nb. tours	25	885	42	41	5	4	9	1,011
	(column %)	(0.93)	(35.11)	(0.69)	(1.24)	(2.21)	(2.44)	(0.07)	(2.32)
	Av. trips	2.0	2.0	2.0	2.8	2.0	1.0	4.4	2.1
BW	Nb. tours	229	155	1,511	2,307	118	45	248	4,613
	(column %)	(7.46)	(14.53)	(26.03)	(30.07)	(9.38)	(36.99)	(7.16)	(16.24)
	Av. trips	4.0	2.0	2.3	2.2	2.1	1.0	4.0	2.6
IM PT	Nb. tours	13	221	8	17	0	1	245	505
	(column %)	(0.58)	(5.44)	(0.04)	(0.67)	(0.00)	(0.02)	(1.47)	(1.04)
	Av. trips	2.9	2.2	2.8	2.2	-	1.0	4.2	3.3
IM BW	Nb. tours	43	42	145	191	10	0	853	1,284
	(column %)	(0.80)	(1.77)	(2.60)	(3.70)	(1.62)	(0.00)	(9.30)	(5.16)
	Av. trips	3.6	2.0	3.4	3.4	2.6	-	5.3	4.7
PT BW	Nb. tours	254	86	197	172	11	10	371	1,101
	(column %)	(23.00)	(15.66)	(9.74)	(10.00)	(2.34)	(17.68)	(17.38)	(14.17)
	Av. trips	2.3	2.2	2.8	2.6	2.0	1.0	4.9	3.6
IM PT BW	Nb. tours	42	14	12	26	2	0	207	303
	(column %)	(2.04)	(2.19)	(0.76)	(1.61)	(0.16)	(0.00)	(5.37)	(2.91)
	Av. trips	2.7	2.5	3.2	2.5	2.0	-	5.5	4.7

- = missing data
Row labels: Individual Means (IM), Public Transport (PT), Bicycle and Walking (BW)
Column labels: Home – Work – Home (HWH), Home – Education – Home (HEH), Home – Personal Activity – Home (HPH), Home – Social Activity – Home (HSH), Home – Transport Someone – Home (HTH), Home – Home (HH), Home – Combination of Activities – Home (HxH)

The mean number of trips composing the tour is an indicator of the tour complexity, which has an influence on mode choice that is often neglected by trip-based models. By definition, it is equal to one for HH tours. Considering first all tours irrespective of travel means (rows 1-3 in Table 1) the mean number of trips is between

2.1 and 2.7 for tours containing only one kind of activity, education-related tours being the least complex and personal duties-related ones the most complex. Obviously, the mean number of trips sharply increases for HxH tours.

Considering the last column of Table 1, the first noticeable aspect is the wide use of individual motorized means, according the typical trend in the U.S. However, public transport plays a non-negligible role, since it is used alone or in combination with other modes in 20.44% of tours taken in New York State (this figure lowering to 6.1% for the whole U.S., see Pirra and Diana (2016), Table 3), while 23.28% of tours involve the use of more than one travel means.

Observing rows 4 onwards of the table, it is possible to notice the intertwined relationship between tour-level mode choice, tour complexity and related activity patterns. For example, the great majority of tours whose purpose is to transport someone are done by individual motorized means (84.29% of tours for HTH). Many students going to school, college or university use public transport, either alone (35.11%) or in combination with other means. On the other hand, walking and biking is much more frequent for HPH, HSH and HH tours.

From this preliminary analysis it is clear that people vary their modal choices in relation with both their activity patterns and the complexity of their tours. On the other hand, it is well known that travel-related choices can also be traced back to the socioeconomic characteristics of individuals. Only this latter group of travel determinants can be easily captured through a trip-level travel demand analysis. In the next section, the goal is to assess if the previously presented classification technique can give a contribution in understanding the relative influence of personal characteristics and contextual or travel related variables in tour-based mode choice.

SVM implementation

A SVM classification technique is implemented to understand which factors can explain at best the observed mode choices at the tour level. Consistently with the prevailing terminology used in this research field, both personal and tour characteristics are called through the term "features".

Table 2. Features used in the classification process. P: personal feature, T: travel-related feature

Feature	Description	Derivation from NHTS Variables	Type
N_AGE	Respondent age	Taken from R_AGE	P
N_INC	Household annual income	Derived from HHFAMINC as mean of income bracket boundaries; for the last income range ≥ $100,000 a weighted value is derived according to the income distribution of the New York State	P
Y_MALE	Respondent gender	Dummy variable from R_SEX: 1 if male, 0 if female	P
Y_URB	Household in an urban area	Dummy variable derived from URBRUR: 1 if household in urban area, 0 if in rural area	P
N_ACT	Number of activities	Number of activities done during the tour	T
Y_EDU	"Education" activity in the tour	Dummy variable: 1 if one of the tour purposes is "Education", 0 otherwise	T
Y_OTH	"Other" activity in the tour	Dummy variable: 1 if one of the tour purposes is "Other", 0 otherwise	T
Y_WORK	"Work" activity in the tour	Dummy variable: 1 if one of the tour purposes is "Work", 0 otherwise	T
Y_WKD	Tour is done on weekday	Dummy variable: 1 if tour is totally on a weekday, 0 if it is at least partially on a weekend	T

In Table 2 the nine features that are going to be considered are listed: the first four are related to personal characteristics (P) and the last five to tour attributes (T). The information available in the dataset is obviously much richer, and the list of know determinants of modal choices much longer: however, the focus here is on a small subset of variables to keep the classification scheme relatively simple and ease the interpretation of the results. Following the research goal stated at the end of the introduction, the relative predictive power of the two groups of determinants in studying modal choices is assessed.

Thus, the main goal of this research is the application of the SVM technique in a multi-dimensional space defined by the variables listed in Table 2. Tours are classified into one of the seven mode use combinations listed in rows of Table 1. The process can be described through the following five steps:

(1) Features selection: choice of which variables of Table 2 are used to create the multi-dimensional space (either all or some of them).

(2) Normalization: features are normalized through a min-max normalization so that each variable stands in a [0 1] range, to provide a value homogeneity with the binary features.

(3) Cross validation process: the dataset is segmented in 5 equal-sized partitions, sampled randomly but maintaining the classes distributions of the whole dataset, and one of these partitions is used as testing (20% of data) while the others are used for training the classifier (80% of data). The procedure is run 5 times, so that each partition is used once for testing (Tan, Steinbach, and Kumar 2005).

(4) SVM classification: the training data are used to create the classes and, then, labels are assigned to the testing part.

(5) Labels comparison: since real classifications are known, it is possible to check the accuracy of the classifier. For each of the seven above mentioned travel modes combinations, the percentages of assignation to the seven different classes are computed.

The above procedure was developed in KNIME, which is an open source platform useful for data analysis (Berthold et al. 2009). It is based on a modular data pipelining approach that allows the connection of different components representing machine learning and data mining techniques. Another interesting characteristic of the platform is the possibility of integrating other open source projects, such as machine learning algorithms from Weka and LibSVM, which were needed given the imbalanced sizes of the classes considered: from Table 1 it is apparent that the biggest group (IM) is 100 times larger than the smallest one (30,350 occurrences against 303 of "IM PT BW").

Classification algorithms require usually a reasonably even distribution of data among the available classes, in order to assure a better domain definition and, thus, better performances. This is achieved if the learner is "trained" on the widest range of elements belonging to each group considered (Tan, Steinbach, and Kumar 2005). However, imbalanced class distributions are rather common in real world applications and various procedures are provided to deal with them, such as undersampling, oversampling, Synthetic Minority Oversampling Technique (SMOTE) or cost sensitive techniques (He and Garcia 2009). These techniques can be used for SVM, but also algorithmic modifications are available to reduce its sensitivity to class imbalance (Batuwita and Palade 2013).

A possible solution, which is the one adopted, is related to the imbalanced support vector ratio. In Wu and Chang (2003) the authors show that, as the training

data gets more imbalanced, the ratio between support vectors belonging to the positive and negative classes (respectively called positive and negative support vectors) becomes more imbalanced too. As they hypothesize, the neighbourhood of a test instance close to the boundary is thus more likely to be dominated by negative support vectors and a boundary point is more likely to be classified as negative by the decision function. Akbani et al. suggest acting on the weights in the decision function, leading to different weights on the negative/positive support vectors (Akbani, Kwek, and Japkowicz 2004). Since this solution can be realized in KNIME through the tool LibSVM (Chang and Lin 2011), this procedure is followed, therefore providing the appropriate weights to each class.

Results and discussion

Results of the classification exercise for different computational experiments are shown in Table 3. The second column of the table indicates if all nine features proposed in Table 2 were considered as dimensions of the space for SVM ("A" rows), or whether the solution space is reduced by taking into account only the five transport-related features ("T" rows), or, alternatively, only the four personal features ("P" rows).

Table 3 comes in the form of a 7*7 matrix for each of the three above analyses, where percentages indicate the fraction of tours of the class that is specified in the first column that is classified into each of the seven classes by the algorithm. Elements lying on the main diagonal of the matrix thus represent a measure of accuracy of the classification, with an identity matrix representing the ideal outcome. Such elements are reported in bold. However, it should be noticed that some of the classes represent a mixed use of different travel means, so that not all mismatches are equally problematic. It is in fact more difficult for example to discriminate tours in which only cars are used from those in which cars are jointly used with other means. Interestingly enough, the

matrix is far from being symmetric. Therefore, mismatches between any two pairs of classes are not probably only due to a lack of discriminating power of the method, but rather to the influence of different factors that will be commented in the following.

Table 3. Row percentages of tours as labelled by SVM.

%		Assigned class						
Real class		IM	PT	BW	IM PT	IM BW	PT BW	IM PT BW
IM	A	**19.51**	1.39	43.02	1.38	15.73	12.68	6.29
	T	**1.77**	1.58	45.38	0.99	34.63	11.97	3.68
	P	**31.98**	4.87	3.97	7.24	21.63	16.72	13.59
PT	A	0.40	**81.60**	7.02	6.43	0.20	3.86	0.49
	T	0.00	**86.75**	8.41	1.48	0.59	2.77	0.00
	P	2.37	**44.81**	0.50	41.15	1.48	7.91	1.78
BW	A	9.30	3.06	**73.32**	0.67	2.47	9.32	1.86
	T	0.56	3.30	**82.03**	0.43	8.00	4.81	0.87
	P	23.00	9.58	**4.34**	7.44	18.25	24.19	13.20
IM PT	A	3.96	36.83	1.98	**43.57**	5.74	3.37	4.55
	T	0.00	42.57	2.18	**39.61**	11.68	1.78	2.18
	P	2.97	33.07	0.79	**55.25**	2.77	2.97	2.18
IM BW	A	13.71	2.88	11.61	5.22	**41.43**	8.72	16.43
	T	1.32	3.27	8.18	4.83	**65.66**	2.18	14.56
	P	24.38	7.55	3.66	9.35	**24.45**	14.72	15.89
PT BW	A	6.81	7.17	23.89	4.36	12.99	**34.88**	9.90
	T	1.18	7.45	25.80	6.27	32.33	**22.16**	4.81
	P	13.08	10.17	3.72	6.36	13.26	**40.60**	12.81
IM PT BW	A	8.58	4.29	9.24	12.87	29.70	14.19	**21.12**
	T	0.66	3.96	7.26	17.16	44.23	11.88	**14.85**
	P	12.87	10.56	1.98	14.19	18.81	19.80	**21.79**

Beyond the three basic “A”, “T” and “P” analyses, additional ones were run in which different mixes of “T” and “P” features were considered. Detailed results are not presented here but will be later recalled whenever they are useful to understand the influence of specific variables on modal choice. The comparison of the results from different analyses gives in fact interesting insights on the relative importance of modal choice determinants, and on how their relative importance changes according to different travel means.

In the following, the outcomes for the four typologies of tours involving the use of only one mode are firstly proposed. Then, the analysis focuses on those referring to the remaining three multimodal classes. The ability of the classifier in properly recognizing the tours done with specific means, or with a combination of them, together with the assignment of a large number of elements to the wrong class, can provide useful and interesting suggestions on the influence that socioeconomics and travel-related features have on mode choice.

Monomodal Classes

Considering tours done only using individual motorized means (first three rows of Table 3), results are generally worse than those related to other classes. A possible explanation stands in the high variability characterizing this class, representing the most commonplace mobility habit in the U.S. and therefore collecting a large number of elements which may not have many common characteristics. However, there is a clear difference between the analysis that only considers personal features and the other two. In the former case, almost 32% of IM tours are correctly classified, while car use is predicted together with other means for an additional 42.46% of cases. On the other hand, when only travel-related features enter into the analysis, IM tours are predominantly classified as tours involving the use of active means (bike and feet). Against intuition, considering an additional travel related feature such as trip distance improved the latter results but only marginally (IM tours that are correctly identified increase from 1.77% to 11.85% and IM tours labelled as BW tours decrease from 45.38% to 38.85%). Therefore, trip distance was excluded from the features listed in Table 2. The problems of recognition in this class are probably due to its large dimension, since it collects more than 77% of all tours (see Table 1), well attesting the widespread use of cars irrespective of the travel patterns by a significant portion of the

population. Many of these tours seem to have transport-related features that would make them suitable for active means; conversely, personal features of motorists are distinct from those of travellers walking and biking (31.98% of correct matching versus 3.97% of incorrect classification of IM tours as BW ones).

On the other hand, IM tours have features quite different from PT tours, so that the substitution potential of private means by public transport seems rather limited considering this tour-level analysis (incorrect classifications of IM tours as PT ones ranging from 1.39% to 4.87%). The latter result would have been hardly observable in a trip-level analysis which does not consider travel constraints at the tour level.

Interesting results are obtained modifying the features space used in the classification. Jointly considering all "T" features and one of the socioeconomics variables, or a combination of them, is not able to ameliorate the results obtained when only "T" features are considered. On the contrary, considering the socioeconomics domain, which is characterized by a rather correct recognition pattern, and adding any "T" feature, results are worsened and they resemble to those reported in the first row of Table 3 except when only "Y_EDU" and/or "Y_WKD" are considered. In the latter case, results in the third row of Table 3 are almost unchanged. All "T" features except those two have therefore a small discriminatory power to single out IM tours.

Tours only involving the use of public transport are well categorized (see bold percentages in rows 4-6 of Table 3). Contrary to IM tours, tour-related features have much higher predictive power than personal characteristics. A deeper analysis of the features reveals that PT tours have the highest occurrence on weekdays ("Y_WKD" variable) and that "N_ACT" is almost always equal to 1. This is due to the fact that the majority of PT tours are for education purposes and are made by underage students

travelling by school bus. Therefore, socioeconomic features alone and especially age already have a good predictive power compared to IM and BW tours.

Tours done by active means show similar patterns concerning the importance of travel-related features, while personal features are not at all good predictors. In other words, the choice of both public transport and bike and walking seems more linked to the considered tour-related patterns, while bikers and walkers are not really distinguishable from motorists through the personal features considered (23.00% of BW tours are classified as IM ones). The latter result seems to contradict that, as stated previously, "personal features of motorists are distinct from those of travellers walking and biking", given the fact that less than 4% of IM tours are classified as BW tours. The explanation lies in the fact that bikers and walkers have a more specific profile that can be shared with that of some of the motorists, while motorists' profiles are much more heterogeneous and they cannot be all assimilated to those of bikers and walkers. The capability of pointing out such asymmetries is one attractive feature of this classification technique, compared to more synthetic goodness of fit measures or significance levels of exogenous variables in statistical inference methods, which can give less punctual indications.

Finally, the check if the addition of a "T" variable would increase the recognition percentage for this class is done. The results show that this happens with "N_ACT" and "Y_WORK", showing that these two tour characteristics play a role in the choice of active means.

Multimodal Classes

In the dataset, multimodal classes represent a smaller number of tours if compared to monomodal ones (see the last column of Table 1). Looking at the corresponding rows in the bottom half of Table 3, the classifier provides usually reasonable matchings, even if

accuracy is sometimes not very high. However, for such multimodal tours SVM tends to assign some of them, usually more than 10%, to a unimodal class implying the use of one of the corresponding travel modes. For example, 43.57% of “IM PT” tours are correctly labelled, while 36.83% of them go in the PT class. This reinforces the previous finding on the relatively tenuous difference between multimodality and monomodality choice in a tour.

Tours combining motorized individual and public transport means are not clearly distinguished from monomodal public transport tours, while they are seldom misclassified as monomodal IM tours. Considering the pervasive use of cars in the sample, the use of public transport is the real distinguishing feature considering both socioeconomic and travel related features.

On the other hand, the joint use of individual motorized modes and active means is best explained when considering travel-related features of tours, while personal characteristics tend to blur the difference with the exclusive use of cars. Conversely, personal characteristics are important to correctly classify “PT BW” tours, since travel features would made them hardly distinguishable from monomodal BW tours.

A more disaggregate analysis shows that only the addition of “N_INC”, alone or in association with other variables, improves the accuracy results related to the class PT BW by avoiding considering them as IM BW tours. When a tour cannot be easily accomplished by simply walking or cycling, income has a determining role in the choice of which additional mode to use.

Finally, more complex tours involving the use of more than two different travel means are obviously more hardly distinguishable by a classification algorithm, especially from other multimodal tours with only two different travel modes.

Conclusion and future work

In this paper Support Vector Machine, a classification technique, has been used to analyse a tour-based travel demand dataset, in order to check its ability in recognizing defined trends in travel mode choice. The research concentrates on the New York State dataset, where tours done during a day are derived from the U.S. National Household Travel Survey administered in 2009. We define from this database, and for each personal tour, nine variables that were then used to build the features space in which the SVM operated its classification process.

When the entire set of these variables is taken into account for the evaluation and only one mode is used in the tour, good recognition results are obtained with the exception of the individual motorized mean class. The explanation stands in the large number of differently characterized people using cars for their daily travels. The public transport class is usually well recognized, since it refers mainly to a specific type of tours, those with education purpose. Thus, the relation with the feature "N_AGE" is certainly rather strong. Travel-related aspects seem to have higher influence in the choice of active means if compared to the socioeconomic ones. In particular, "N_ACT" and "Y_WORK" helps in the proper class recognition.

In general, a smaller number of tours in the dataset is multimodal. In such cases, multimodal classes are often labelled as one of the related unimodal ones. Results show in particular how income has a determining role in the choice of using either public transport or an individual motorized means to a tour where the traveller is already biking or walking.

The experience gained in applying SVM to study mode choices at the tour level shows the potential complementarities between such classification approach and more popular statistical inference and econometric models. In particular, SVM implementation is relatively easy and straightforward and it can predict to some extent,

in a real settings where car use dominates, which tours are likely to be made by public transport or non-motorized means. Additionally, it is possible to assess how the predictive power of each feature changes according to the combination of travel means used in different tours, so that the method can be adapted to the specific problem under consideration. As an example, considering the number of activities in a tour beyond a set of socioeconomic characteristics of the traveller improves the recognition of tours involving the use of public transport and/or active means, while it worsens the recognition of tours partially or totally travelled by car. Such kind of assessment is more detailed compared to a measure of fit or significance level associated to an exogenous variable in a statistical model.

To summarize, while the accuracy of results is probably not comparable to that of a good model, SVM can give a first approximation answer in case studies were large amount of data need to be quickly processed and heuristic solutions are acceptable. Potential applications could range from activity-based travel choices simulators to search engines supporting personalized travel planners. Also on a policy viewpoint this technique can be useful. By considering mismatches in a classification exercise, the analyst can identify which monomodal car tours could be completed by active means (these latter are IM tours that are classified as BW ones). These tours can then be the specific target of marketing campaigns such as voluntary travel behaviour change programs.

A new method to study modal choice, based on a classification technique, has been proposed. Future work will concentrate on checking the behaviour of the classifier for more targeted kinds of tours, based on the related activity patterns, and in developing better classifying features beyond the consideration of basic travel-related or socioeconomic characteristics of the traveller. Moreover, given the innovative aspect of

the proposed approach, another interesting development is to compare the obtained results with those derived from traditional mode choice models.

References

Akbani, Rehan, Stephen Kwek, and Nathalie Japkowicz. 2004. "Applying Support Vector Machines to Imbalanced Datasets." *Lecture Notes in Computer Science* 3201: 39–50. doi:10.1007/978-3-540-30115-8_7.

Axhausen, Kay Werner. 2008. "Definition of Movement and Activity for Transport Modeling." In *Handbook of Transportation Modeling*, 2nd ed., 329–344. Emerald Publishing Group. doi:10.3929/ethz-a-005278091.

Batuwita, Rukshan, and Vasile Palade. 2013. "Class Imbalance Learning Methods for Support Vector Machines." In *Imbalanced Learning: Foundations, Algorithms, Applications*, 83–96. doi:10.1002/9781118646106.

Beelen, Marjan, Isabel Thomas, and Ann Verhetsel. 2005. "Commuting and Urban Structures in Belgium:An Exploratory Data Mining Analysis on 2001 Data." *Proceedings Of The Bivec-Gibet Transport Research Day*, 387–395.

Berthold, Michael R., Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel, and Bernd Wiswedel. 2009. "KNIME - the Konstanz Information Miner." In *Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization*, edited by C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker. Springer. doi:10.1145/1656274.1656280.

Burges, Christopher J C. 1998. "A Tutorial on Support Vector Machines for Pattern Recognition." *Data Mining and Knowledge Discovery* 2 (2): 121–167. doi:10.1023/9715923555.

Chang, Chih-Chung, and Chih-Jen Lin. 2011. "LIBSVM: A Library for Support Vector

Machines." *ACM Transactions on Intelligent Systems and Technology* 2 (3): 1–27. doi:10.1145/1961189.1961199.

Cheu, Ruey, Jianxin Xu, Alvina Kek, Wei Lim, and Way Chen. 2006. "Forecasting Shared-Use Vehicle Trips with Neural Networks and Support Vector Machines." *Transportation Research Record: Journal of the Transportation Research Board* 1968: 40–46. doi:10.3141/1968-05.

Daly, Andrew, and Stanley Zachary. 1978. "Improved Multiple Choice Models." In *Determinants of Travel Choice*, edited by D.A. Hensher and M.Q. Dalvi, 337–357. Saxon House, Sussex.

He, Haibo, and Edwardo A. Garcia. 2009. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 21 (9): 1263–1284. doi:10.1109/TKDE.2008.239.

Hensher, David A., and Tu T. Ton. 2000. "A Comparison of the Predictive Potential of Artificial Neural Networks and Nested Logit Models for Commuter Mode Choice." *Transportation Research Part E: Logistics and Transportation Review* 36 (3): 155–172. doi:10.1016/S1366-5545(99)00030-7.

Hsu, Chih Wei, and Chih Jen Lin. 2002. "A Comparison of Methods for Multiclass Support Vector Machines." *IEEE Transactions on Neural Networks* 13 (2): 415–425. doi:10.1109/72.991427.

Islam, Md Tazul, and Khandker M.Nurul Habib. 2012. "Unraveling the Relationship between Trip Chaining and Mode Choice: Evidence from a Multi-Week Travel Diary." *Transportation Planning and Technology* 35 (4): 409–426. doi:10.1080/03081060.2012.680812.

Kim, Daehyon, Seungjae Lee, and Seongkil Cho. 2007. "Input Vector Normalization Methods in Support Vector Machines for Automatic Incident Detection."

Transportation Planning and Technology 30 (6): 593–608. doi:10.1080/03081060701698235.

McFadden, Daniel. 1973. "Conditional Logit Analysis of Qualitative Choice Behavior." *Frontiers in Econometrics*. New York, USA: Academic Press. doi:10.1108/eb028592.

Nijkamp, Peter, Aura Reggiani, and Tommaso Tritapepe. 1996. "Modelling Inter-Urban Transport Flows in Italy: A Comparison between Neural Network Analysis and Logit Analysis." *Transportation Research Part C: Emerging Technologies* 4 (6): 323–338. doi:10.1016/S0968-090X(96)00017-4.

Omrani, Hichem. 2015. "Predicting Travel Mode of Individuals by Machine Learning." *Transportation Research Procedia* 10: 840–849. doi:10.1016/j.trpro.2015.09.037.

Omrani, Hichem, Omar Charif, Philippe Gerber, and Anjali Awasthi. 2013. "Prediction of Individual Travel Mode with Evidential Neural Network Model." *Transportation Research Record: Journal of the Transportation Research Board* 2399: 1–8. doi:10.3141/2399-01.

Pirra, Miriam, and Marco Diana. 2016. "Classification of Tours in the U.S. National Household Travel Survey through Clustering Techniques." *Journal of Transportation Engineering* 142 (6): 1–13. doi:10.1061/(ASCE)TE.1943-5436.0000845.

Seyedabrishami, Seyedehsan, and Yousef Shafahi. 2013. "A Joint Model of Destination and Mode Choice for Urban Trips: A Disaggregate Approach." *Transportation Planning and Technology* 36 (8): 703–721. doi:10.1080/03081060.2013.851507.

Shukla, Nagesh, Jun Ma, Rohan Wickramasuriya, and Nam N. Huynh. 2013. "Data-Driven Modelling and Analysis of Household Travel Mode Choice." In *20th International Congress on Modelling and Simulation, Adelaide, Australia*, 92–98.

Shukla, Nagesh, Jun Ma, Rohan Wickramasuriya, Nam N. Huynh, and Pascal Perez. 2015. “Tour-Based Travel Mode Choice Estimation Based on Data Mining and Fuzzy Techniques.” In *International Symposium for Next Generation Infrastructure, United Kingdom*, edited by T. Dolan and B. S. Collins, 215–220.

Tan, Pang-Ning., Michael. Steinbach, and Vipin Kumar. 2005. *Introduction to Data Mining*. Pearson Addison Wesley.

U.S. Department of Transportation. 2009. “2009 National Household Travel Survey.” http://nhts.ornl.gov.

Vanajakshi, Lelitha, and Laurence R. Rilett. 2007. “Support Vector Machine Technique for the Short Term Prediction of Travel Time.” In *IEEE Intelligent Vehicles Symposium, Instanbul, Turkey*, 600–605. doi:10.1109/IVS.2007.4290181.

Vapnik, Vladimir Naumovich. 1982. *Estimation of Dependences Based on Empirical Data*. New York, USA: Springer-Verlag.

Wu, Chun Hsin, Jan Ming Ho, and D. T. Lee. 2004. “Travel-Time Prediction with Support Vector Regression.” *IEEE Transactions on Intelligent Transportation Systems* 5 (4): 276–281. doi:10.1109/TITS.2004.837813.

Wu, Gang, and Edward Y. EY Chang. 2003. “Class-Boundary Alignment for Imbalanced Dataset Learning.” *The Twentieth International Conference on Machine Learning (ICML), Workshop on Imbalanced Data Sets, Washington DC*, 1–8.

Xian-Yu, Jian-Chuan. 2011. “Travel Mode Choice Analysis Using Support Vector Machines.” In *11th International Conference of Chinese Transportation Professionals (ICCTP), Nanjing, China*, 360–371. doi:10.1061/41186(421)37.

Xie, Chi, Jinyang Lu, and Emily Parkany. 2003. “Work Travel Mode Choice Modeling with Data Mining Decision Trees and Neural Networks.” *Transportation Research*

Record: Journal of the Transportation Research Board 1854: 50–61. doi:10.3141/1854-06.

Yang, Yang, Enjian Yao, Hao Yue, and Yuhuan Liu. 2010. "Trip Chain's Activity Type Recognition Based on Support Vector Machine." *Journal of Transportation Systems Engineering and Information Technology* 10 (6): 70–75. doi:10.1016/S1570-6672(09)60073-8.

Zhang, Yunlong, and Yuanchang Xie. 2008. "Travel Mode Choice Modeling with Support Vector Machines." *Transportation Research Record: Journal of the Transportation Research Board* 2076: 141–150. doi:10.3141/2076-16.