ISIM 2011, HEC Montreal

Evidence Theory-Based Mode Choice Modeling

Anjali Awasthi, Hichem Omrani, Omar Charif, Philippe Trigano

CIISE, Concordia University, Montreal
CEPS/INSTEAD research centre, Luxembourg

Travel mode choice problem

- The travel mode choice is a pattern recognition problem (supervised learning), in which several variables (e.g., human characteristics and geographical patterns) explain the choices among the modes (private car, bus, train, bike or foot). We assess models and estimation procedures by the quality of the prediction based on them.

- We apply ENN for predicting the travel mode.

- The modes considered are private car, public transport (bus or train) and soft mode (walking or cycling).

Literature Review

TABLE I
METHODS AND LITERATURE FOR MODELING TRAVEL MODE

Model	Benchmark method	Application scope	Model topology	Interpretability	Computation time	Validation L–T	Optimisation algorithm
DC	MNL	PR	Layer structure	Explicit utility function	Moderate	(1,1)	ML
	ML	—	—	—	—	—	—
	NL	—	—	—	—	—	—
	MNLwUH	—	—	—	—	—	SML
ANN	MLP	PR and CL	Layer structure	Implicit	Low	(2,1)	BP and GD
	RFB	—	—	—	—	—	—
DT	CART	PR and CL	Tree structure	Explicit	Fast	(2,1)	RP
SVM	Linear	CL	CBR	Explicit	Fast	(2,1)	SRM/MCM
SVM	RBF	—	—	—	—	—	—
k-NN		CL	CBR	Explicit	Moderate	(2,1)	CR
Bayes	BC	PR and CL	CBR	Explicit	Moderate	(2,1)	ML
Bayes	BBN	—	—	—	—	—	—

Notes: See tables VIII and IX for acronyms.

Cross-validation

- The standard way of assessing the quality of prediction is by splitting the sample into a learning and a testing dataset, denoted by L and T. The model is fitted on L and its performance is evaluated by comparing the fit with the observed values on T.

- In cross-validation, the sample is split into K subsamples, and a random subset of these subsamples forms L and the remainder forms T. Several random splits L and T are drawn, and prediction is evaluated on each pair L–T.

- We denote this method by L–T(K;R), where K is the number of subsets (folds) and R the number of replications. The standard approach is L–T(2,1); L–T(1,1) corresponds to learning and assessment on the entire dataset, without splitting it. For validation purposes, our cross validation technique uses(L–T(K;R), with K = 10 and R =100).

Daily mobility in Luxembourg

TABLE II
MODAL SPLIT, TRAVELING TIME, DISTANCES RUNNING FROM RESIDENCE TO WORK PLACES

	Modal split	Median time	Distance running
	(%)	(min)	(km)
Car only	76	20	15
Car + other modes	2	45	25
Public transportation	13	30	12
Soft mode (walk, bike)	9	5	1
Total	100	20	12

Source: PSELL-3/2007, CEPS/INSTEAD, STATEC

Car-ownership data

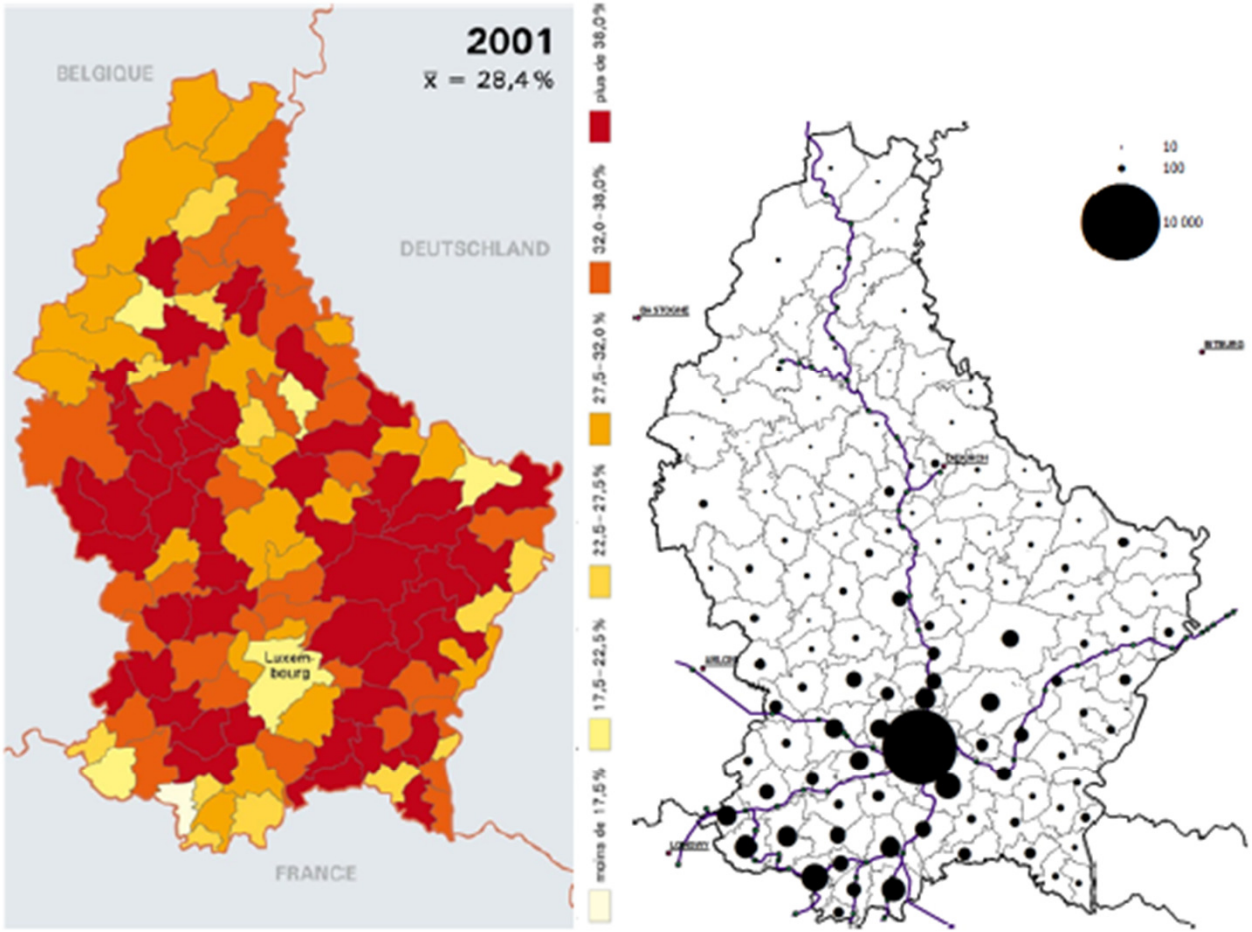

Fig. 1. (a) Proportion of households that have at least two cars; (b) number of workers in the city of Luxembourg. Based on mapping data from the Census 2001.

Modal share

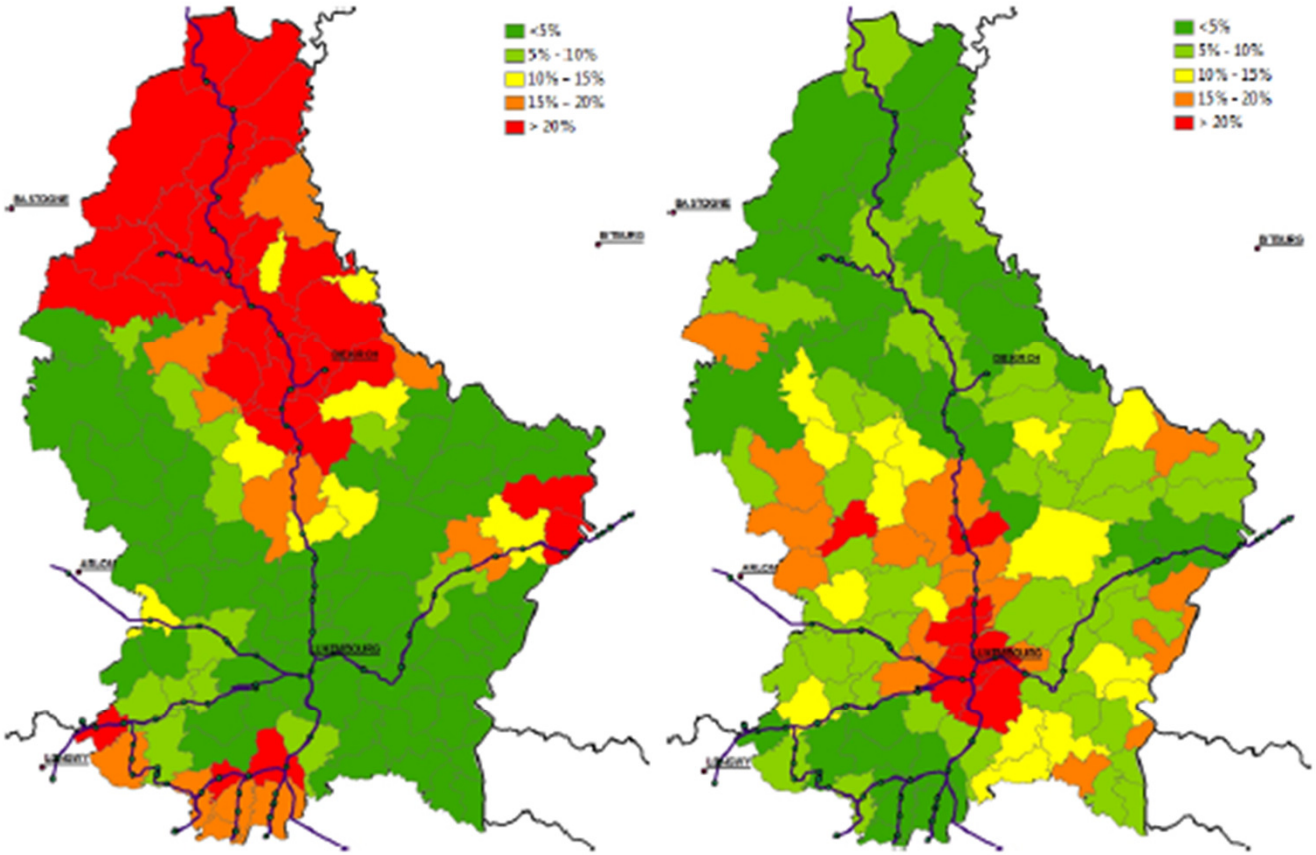

Fig. 2. (a) Train mode share of economically active people working in Luxembourg City; (b) public transport mode shares (Census, 2001).

Subset variable selection

TABLE III
SUB-SET SELECTION

Variable	Number	−Log(likelihood)
All	12	−1432.208
All/{age}	11	−1434.028
All/{age and educational level}	10	−1437.114

We applied backward selection algorithm [39], for subset variable selection. a subset of 10 variables is selected which is considered as the most important variables.

Evidential neural network

- The prediction of the travel mode from explanatory variables is a nonlinear regression problem, because we use the logit link for multinomial outcomes and fit models by generalized least squares.

- The most widely used ANN model is multilayer perceptron network (MLP). We apply ENN with one hidden layer of softmax units trained by minimization of the mean squared error function. This approach is known to provide estimates of the conditional average of the output variable (here, the travel mode choice) given the observed values of the input variables.

ENN architecture

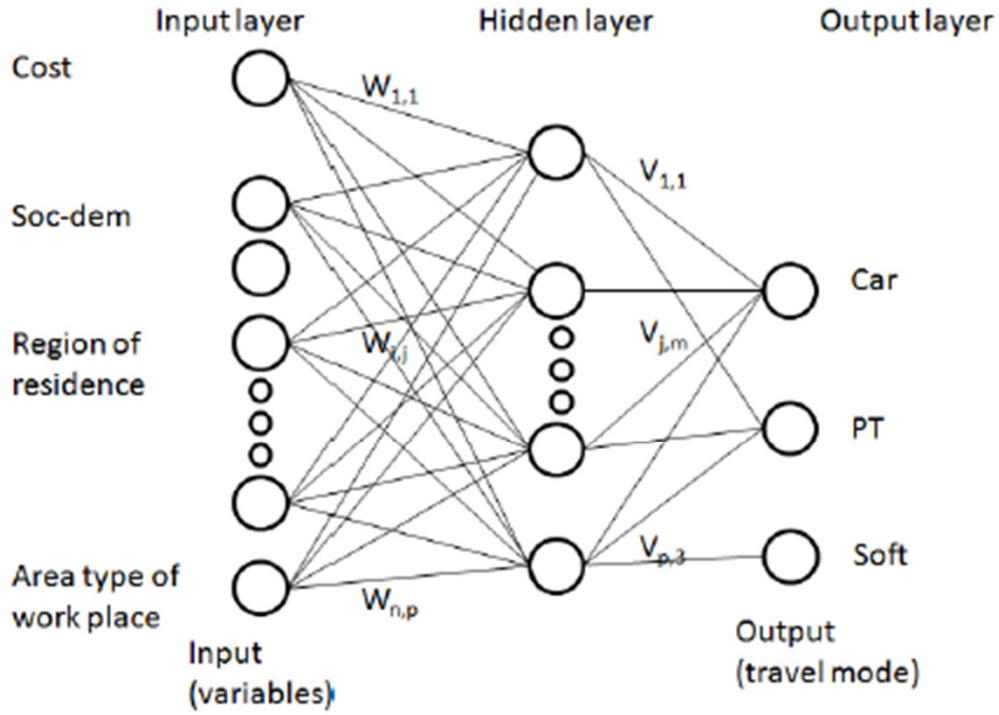

Fig. 3. Topology of the ENN model: MLP applied to predicting travel mode of commuters (soc-dem — social and demographical variables, such as gender, age and nationality).

1 input layer, 1 hidden layer with h = 12 hidden units and 1 output layer

Success rate

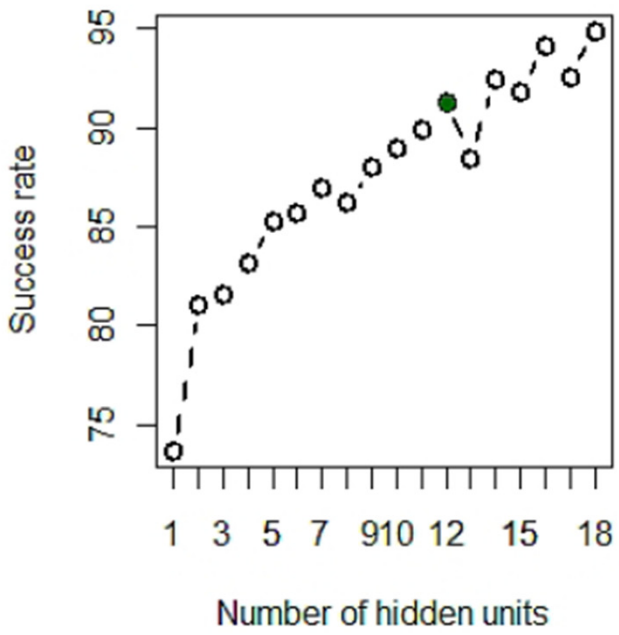

Fig. 4. Proportion of correct predictions (success rate) for number of units in the hidden layer, h, with the optimal solution highlighted.

The number of units in the hidden layer is found by comparing the rates of correct prediction for several numbers h. Using this basis, the optimal choice of h = 12 is highlighted

Case Study

- We use data from the Socio-economic Panel Survey Liewen zu Letzebuerg (PSELL-3). This survey was launched in 2003 with a representative sample of the resident population in Luxembourg.

- The sample size of the survey is around 3500 households (9500 individuals), which allows estimation of social, demographic and economic indicators for the whole population.

- The Survey is carried out annually by the International Network for Studies in Technology, Environment, Alternatives and Development (CEPS/INSTEAD) in collaboration with the Luxembourg Statistical Office (STATEC). It forms the Luxembourg's contribution to the European Union Statistics on Income and Living Conditions (EU-SILC).

- For more details of PSELL-3, see www.ceps.lu/vivre.

Data source and variables

TABLE IV
SUMMARY OF THE MODAL SHARES IN THE DATASET (PSELL-3)

	Total data set		Training data set		Test data set	
Mode	#	%	#	%	#	%
Car	2910	73.7	1749	73.8	1161	73.5
PT	652	16.5	388	16.4	264	16.7
Soft	387	9.8	232	9.8	155	9.8
Sum	3949	100	2369	100	1580	100

The dataset extracted from the PSELL-3 database is composed of 3949 observations, 14 variables

TABLE VII
EXPLANATORY VARIABLES AND DESCRIPTION

Variables	Description
Total travel time (in minutes)	1 to 14.99 min, 15 to 29.99 min, 30 to 59.99 min, 60 min and more
Number of train stations	0 train station, 1 train station, 2 train stations and more (in the municipality of residence)
Number of bus stations (in municipality of residence)	Less than 10 bus stations, 10 to 19 bus stations, 20 to 39 bus stations, more than 40 bus stations
Number of cars in the household	Without car, 1 car, 2 and having more than 2 cars
Age of household	Less than 30 years, 30 to 39, 40 to 49, 50 years and over
Standard of Living	Quartile of equivalised income, with the following thresholds: 24946.10, 33808.90, 45099.50 and 45099.50 €
Household type	Single without children, couple without children, single with children, couple with children
Education	Primary, high school, university, higher non-university degree
Travel distance (CAR_dist) (from home to work in km)	Less than 5 km; 5 km to 9.99 km; 10 km to 14.99 km; 15 km to 24.99 km; more than 25 km
Regions of workplaces	Centre north, centre south (without Luxembourg city), east region, north region, west region, south region, Luxembourg city
Nationality	Luxembourgish, Portuguese, other EU, non-EU
Typology of residence	Dense city, first-ring suburb, second-ring suburb, distant peri-urban, mining area, rural
Gender	Binary (male, female)
Travel cost by car (CAR_cost)	Less than 12.20€ / 3km, 12.20€ / 3km to 22.90€ / 18km, 23€ / 18km to 46.20€ / 50km, more than 46.20€ / 50km
Travel cost by PT (PT_cost)	Less than 22.59€ / 3km, 22.59€ / 3km to 23.05€ / 18km, 23.05€ / 18km to 23.30€ / 50km, more than 23.30€ / 50km
Travel cost by walk (WK_cost)	Less than 12€ / 3km, 12€ / 3km to 24€ / 6km, more than 24€ / 6km
TC travel time (TC_time)	Travel time from home to work by public transport
Car travel time (CAR_time)	Travel time from home to work by private car
Weight (ω^*)	Individual weight in the PSELL-3 database.

Notes: TC_time and CAR_time are computed from Google maps; see Sylvain *et al.* (2010) for details.

CAR_cost = 10 + (CAR_time × 0.183) + (CAR_dist × 0.48) where CAR_time is the declared car travel time and CAR_dist is the declared distance using car.

PT_cost = $20 + 2.143 + 0.583 \times \log_{10}(\texttt{TC\_time})$ where TC_time is the declared time using PT.

WK_cost = $0.333 \times \frac{\texttt{WK\_time}}{0.08333}$ where WK_time is the walking time. For more details about computing the value of travel time, see [41].

* — is used to highlight that the modeling framework with the set of exploratory variables, takes into account the sample weighting (noted by ω).

Geographical information

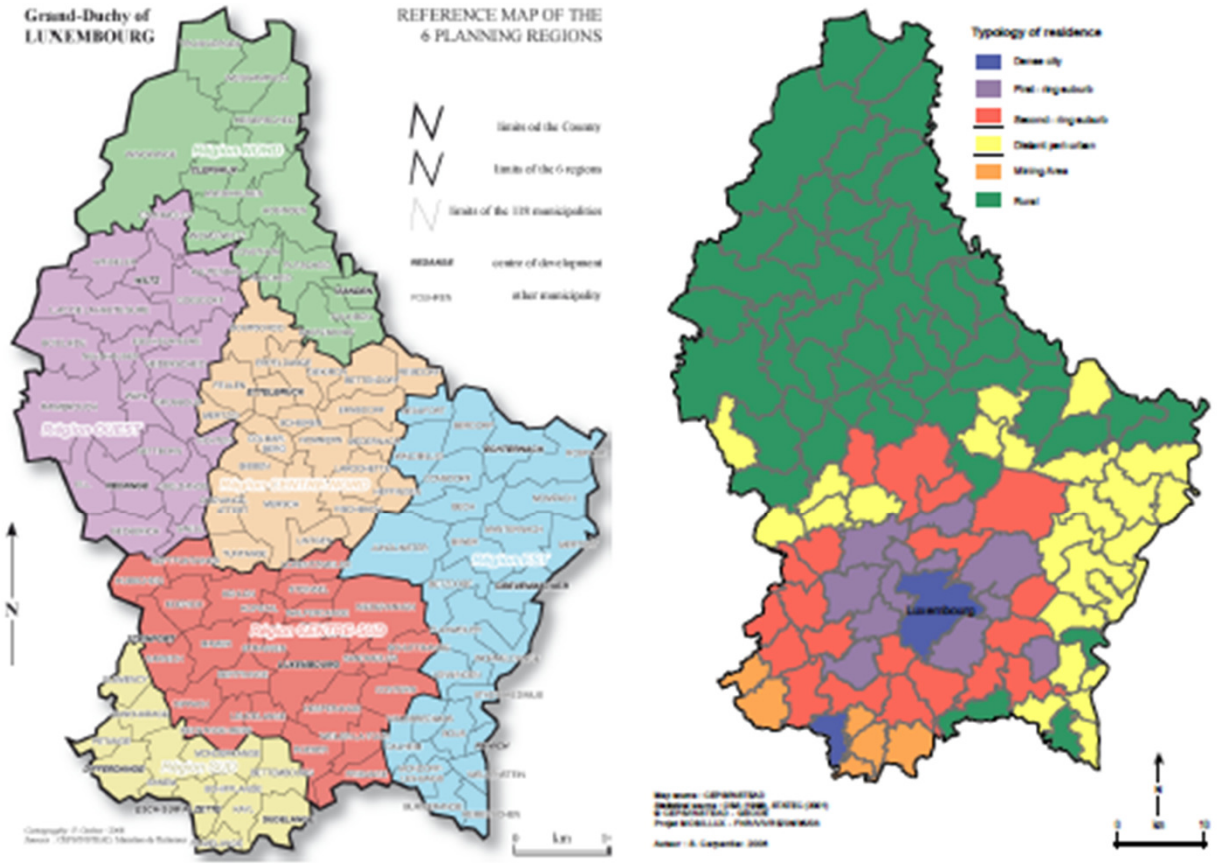

Fig. 5. (a) Regions of Luxembourg; (b) regions and municipalities of Luxembourg. Source: Carpentier (2010).

Input and Output Variables

- The input variables are classified to the following groups (dimensions).
  - C: cost
  - D: income, age, gender, nationality, type of household, education
  - T: car ownership, number of bus stops and train stations in the municipality of residence
  - G: region of residence and area type of work place

- The outcome variable is the travel mode; it has three categories, private car, public transport (PT) and soft mode. We combine walking and cycling mode because their frequency is only 9%;

ENN Model Results

The observed composition of the modes is (73:7; 16:5; 9:8)%.

The result from one replication by splitting the overall sample S to subsamples L and T is given below:

TABLE V
CLASSIFICATION FROM ENN MODEL

Observed mode	Predicted mode			
	Car	PT	Soft	Success rate (%)
Car	2732	97	81	93.9
PT	348	266	38	40.8
Soft	161	22	204	52.7
Overall (%)	82.1	9.7	8.2	83

Predictions – ternary plot

Each individual is represented by a point. Vertices C1, C2 and C3 correspond to certainty that the individual belongs to the respective component 1 (car), 2 (pt) and 3 (soft).

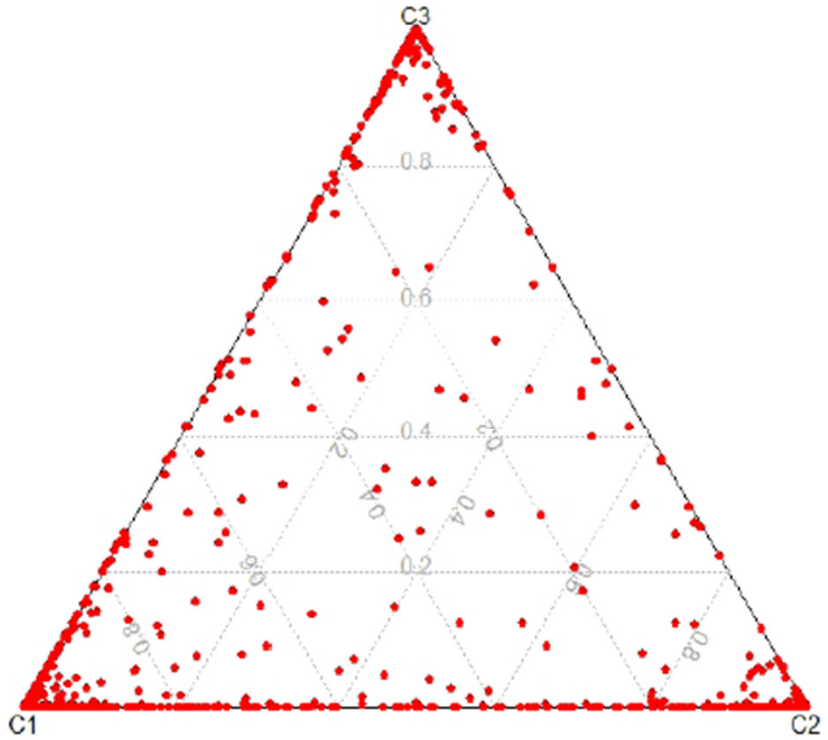

Fig. 6. The ternary plot of the estimated probabilities of belonging to the travel mode; three modes: car, pt and soft.

Confusion matrix and success rates

TABLE VI
PERFORMANCE MODELS (SUCCESS RATE) USING CROSS-VALIDATION

Order	Models †	Success rate	Rank
1	ANN-MLP, $h^* = 5$ units in hidden layer	82	2
2	ANN-RBF, $\gamma^* = 1$	81	3
3	DT: classification tree	78	7
4	Bayes	67	8
5	MNL	62	9
6	k-NN, $k^* = 18$	77	6
7	SVM, kernel RBF, width $\gamma^* = 0.02$	80	4
8	SVM, kernel polynomial, degree $d^* = 3$	79	5
9	ENN, $h^* = 12$ units in hidden layer	83	1

Notes: † — see Appendix B for abbreviations. An asterisk * is used to highlight the optimal value used for the appropriate model.

Each learning algorithm was run 100 times in each configuration.

R programming language

Average test error rates and standard deviations

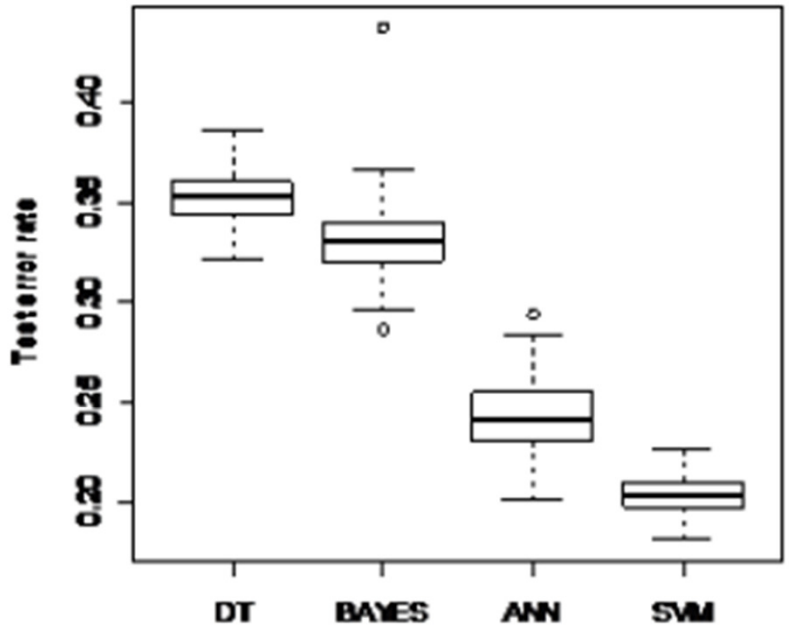

Fig. 7. Boxplot of the success rate using: DT, Bayes, ANN and SVM,... models

Appendix

TABLE VIII
MODELS AND BENCHMARK METHODS

Acronym	Description	Acronym	Description
ANN	Artificial neural network	ML	Mixed multinomial logit
Bayes	Bayes classifier	MNL	Multinomial logit
BBN	Bayesian belief networks	MNLwUH	Multinomial logit with unobserved heterogeneity
BC	Bayes classifier	MLP	Multi-layer-perceptron
CART	Classification and regression trees	NL	Nested multinomial logit
DC	Discrete choice	*k*-NN	k-nearest neighbor algorithm
DT	Decision tree	RBF	Radial basis function
DST	Dempster-Shafer theory	SM	Softmax
ENN	Evidential neural network	SVM	Support vector machine

TABLE IX
VALIDATION AND OPTIMIZATION METHODS

Acronym	Description	Acronym	Description
BP	Back propagation	ML	Maximum likelihood
CBR	Case-based reasoning	PR	Prediction
CL	Classification	ROC	Receiving operating characteristics
CR	Case ranking	RP	Recurcive partitioning
CV	Cross-validation	SML	Simulated maximum likelihood
GD	Gradient descent	SRM	Structural Risk Minimization
L	Training set	T	Testing set
MCM	Maximization Classification Margin