ALBATROSS Model Development

Outline

- The process model
- Decision tree induction method
- Choice of attribute variables and constraints
- Making decisions in prediction stage
- Data bases
- Decision tree induction results
- Interpreting decision trees
- Conclusions

The Process Model (1)

- A priority-based scheduling process
  - Priority ranking of choice facets
  - Priority ranking of activities
- Qualitative decisions are made as much as possible explicit
- Schedules of household members co-evolve by alternating decisions between their schedules

The Process Model (2)

- Maximally two adult members per household included
- Schedule starts at 3 AM and ends at 3 AM the next day

The Process Model (3)

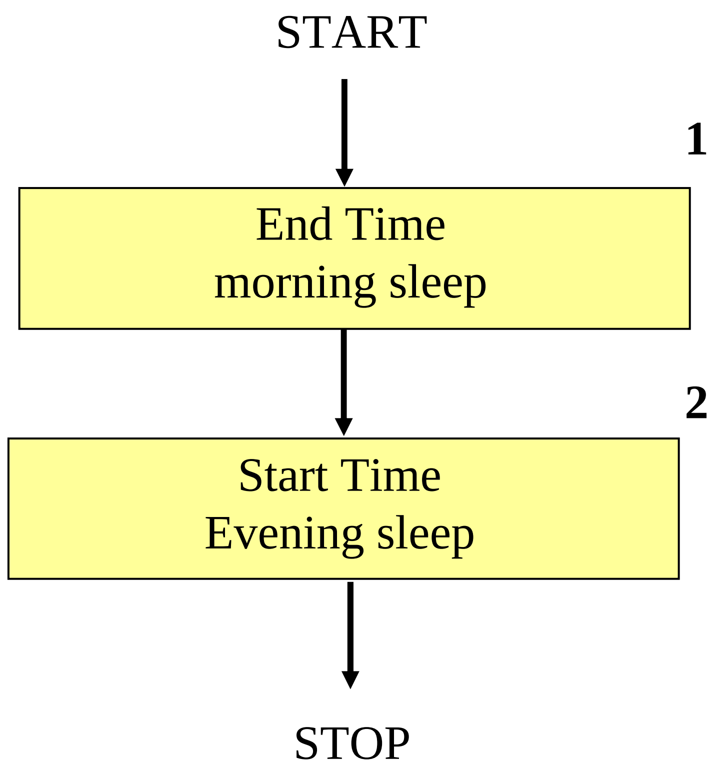
1. Schedule Skeleton

    - Sleep pattern
    - Work/school pattern
    - Secondary fixed activities

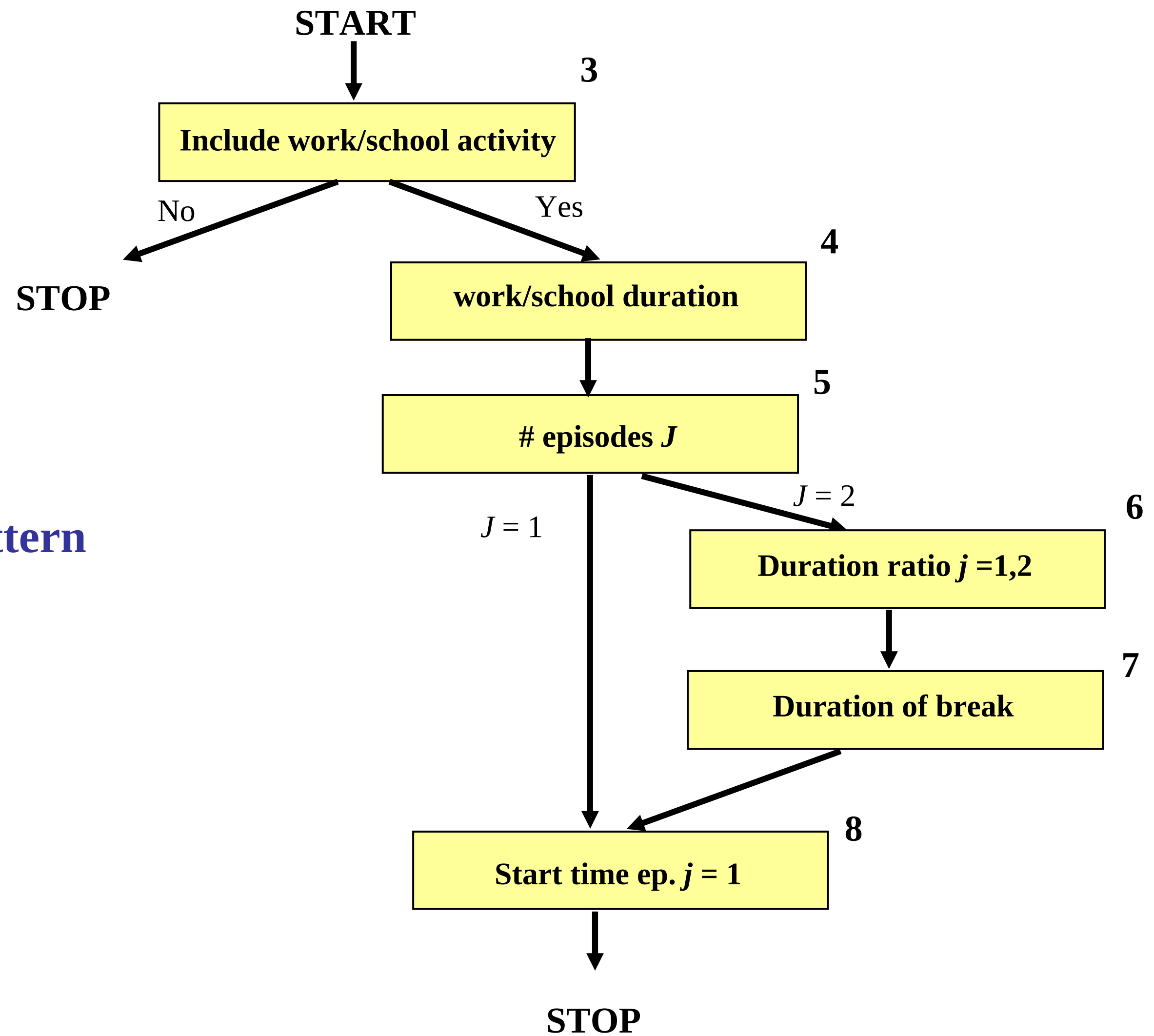
2. Transport mode for work/school trips

3. Flexible activities

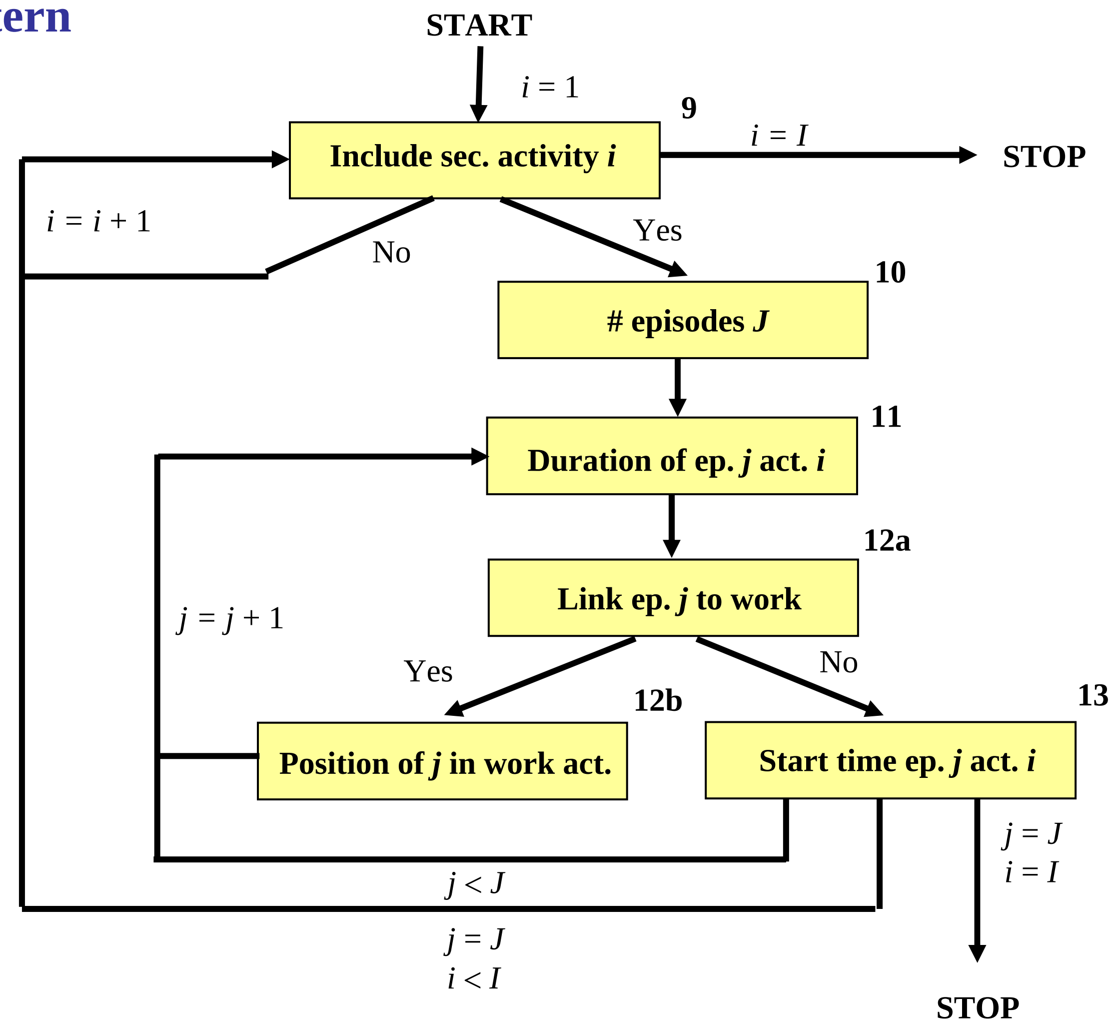
The sleep pattern



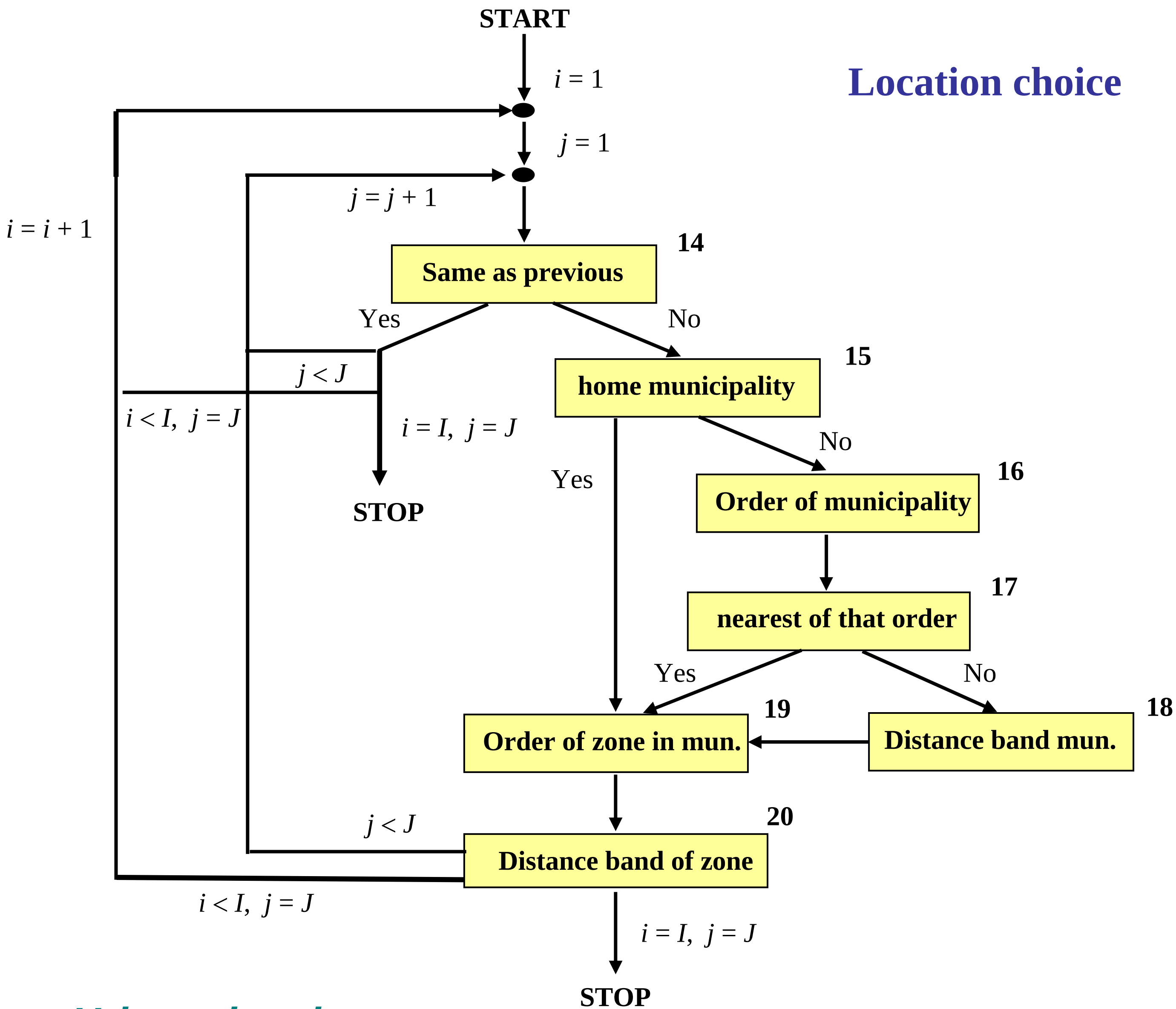
The work pattern



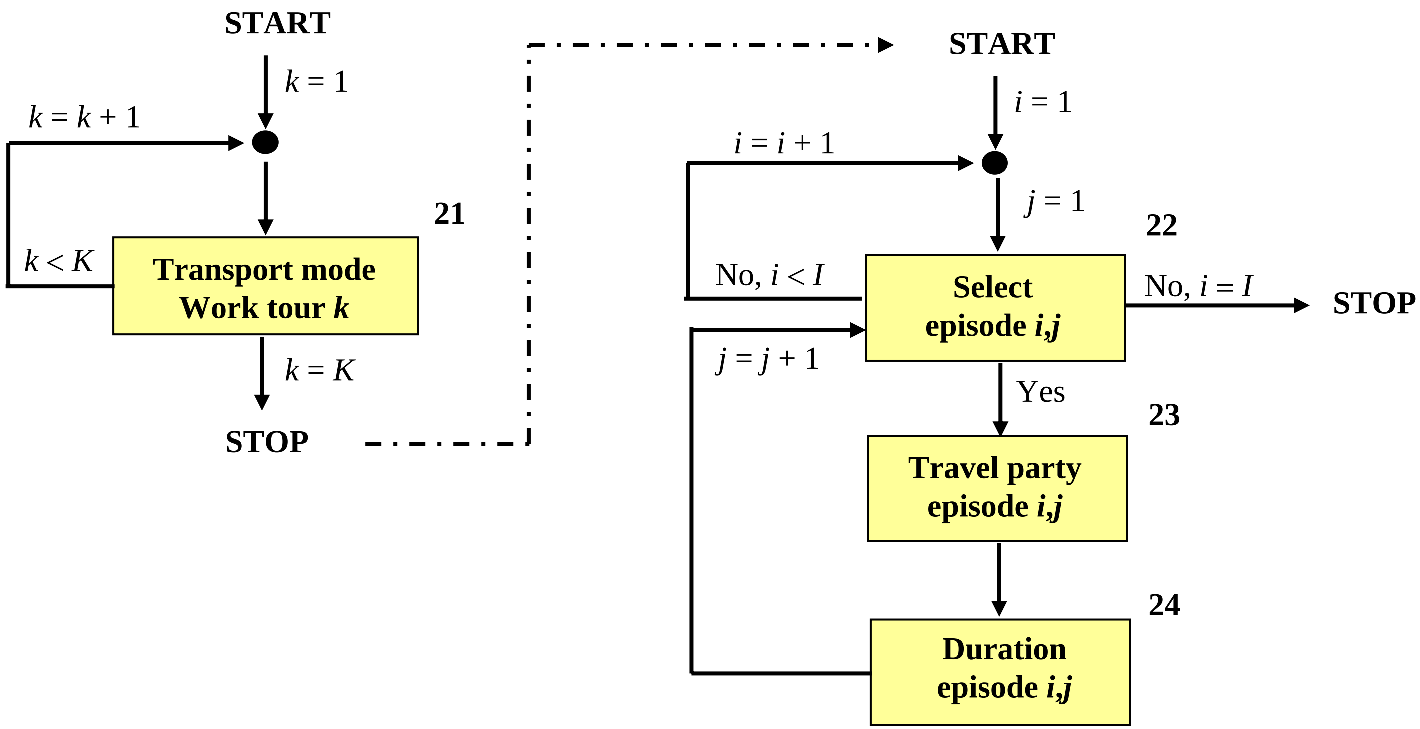
The secondary fixed activity pattern



Location choice

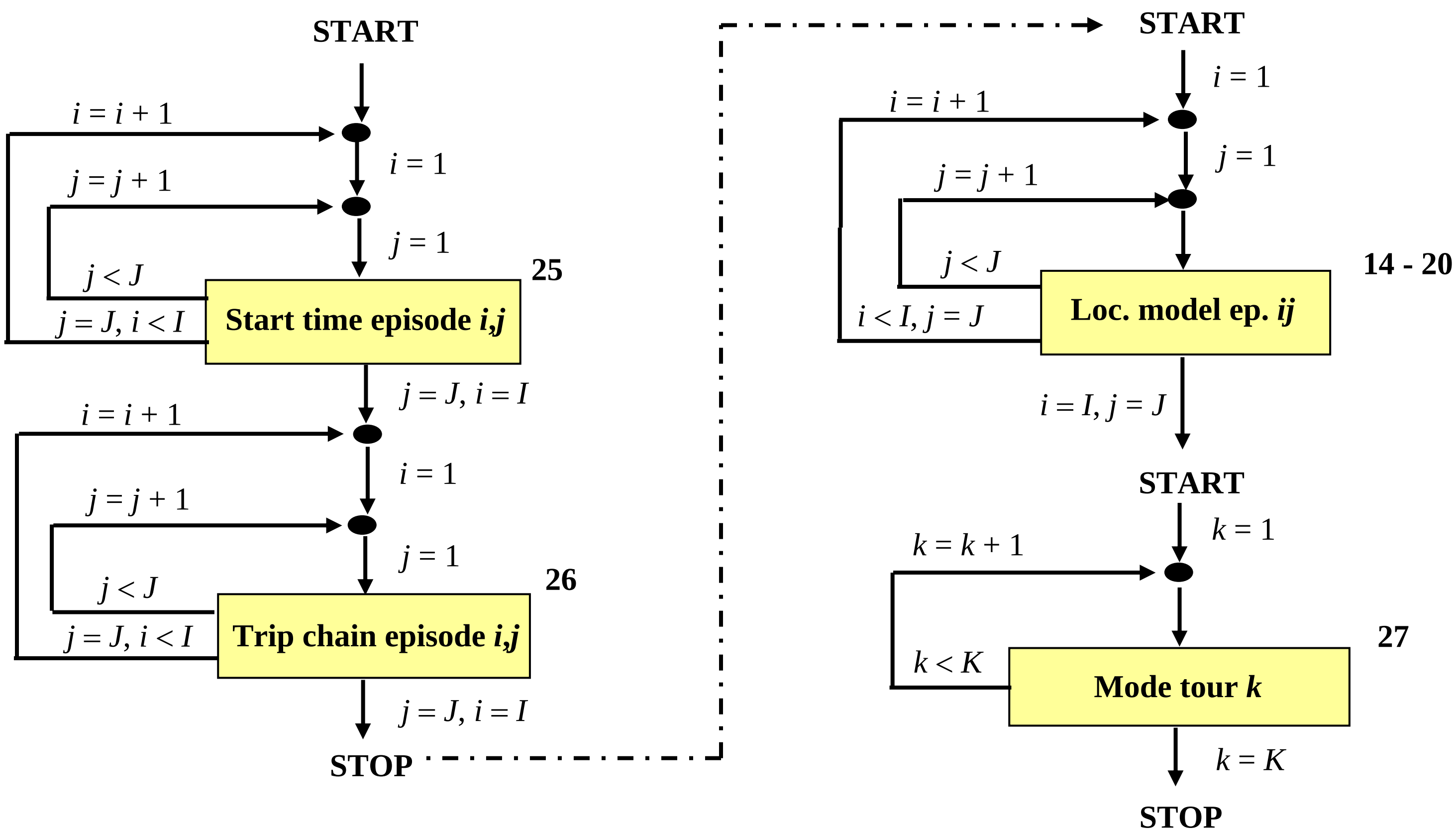


Transport mode to work Flexible activities program



Time and trip chain flexible

Location and mode flexible



Resulting patterns

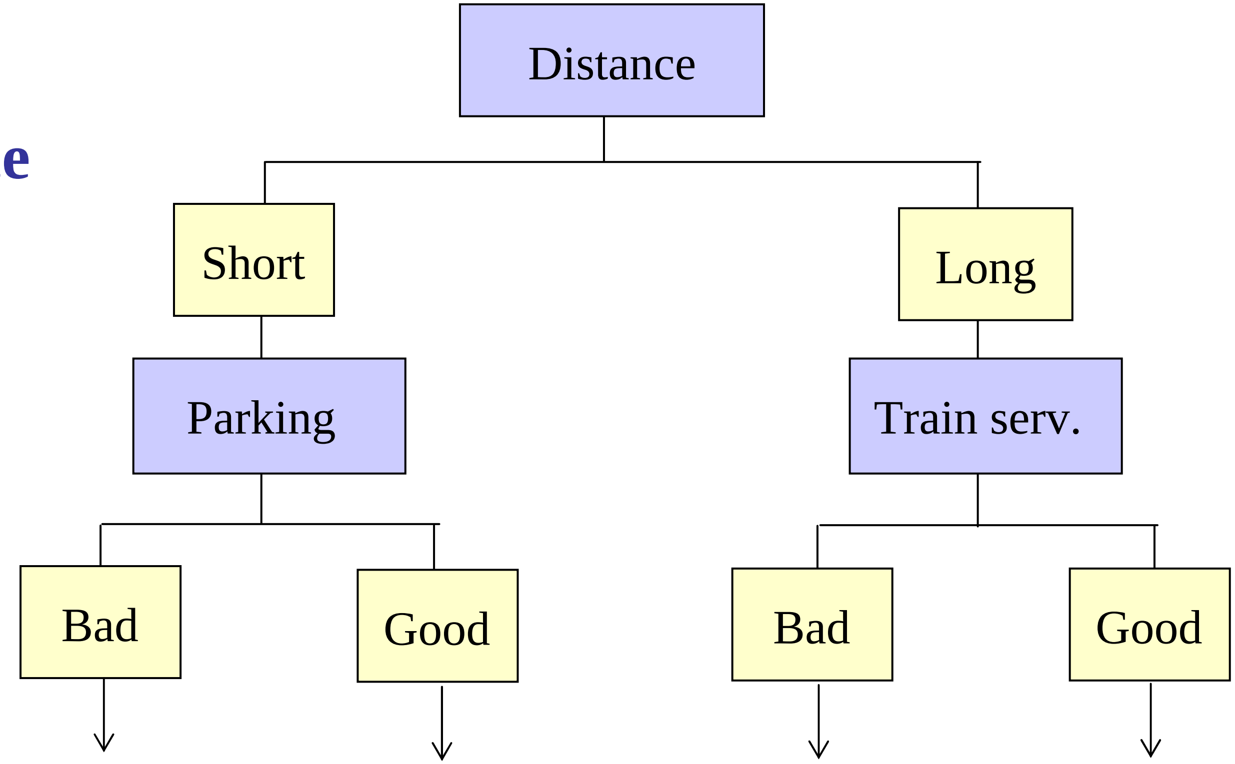
- First and last activity is sleep
- Tours are identifiable
- Number of activities/trips within tours is not restricted
- No mode switches within tours
- Constraints are not violated

Decision tree induction

- Observations are taken from diary data
  - Attributes: $X_{i1}, X_{i2}, \ldots, X_{in}$ for $i = 1 \ldots J$
  - Choice: $Y_i \in \{ 1, 2, \ldots, p \}$ for $i = 1 \ldots J$
- A CHAID-based method recursively splits the sample on *X* into increasingly homogeneous partitions in terms of *Y*
- Significance level is used as a split criterion
  - Chi-square for *discrete* choices
  - F-statistic for *continuous* choices

Decision tree example

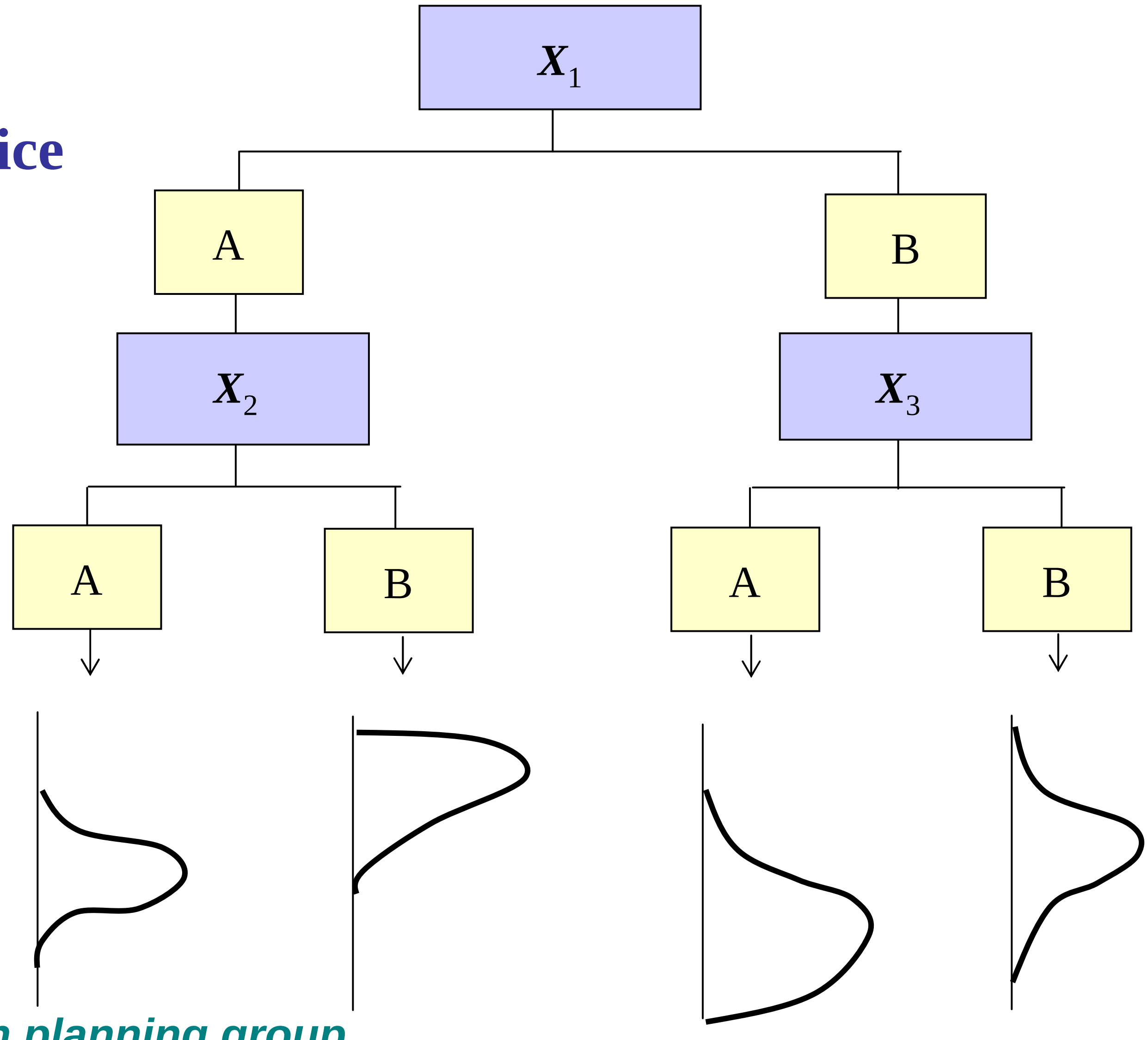
Discrete choice



Slow	0.90	0.71	0	0
Car driv.	0.08	0.29	0.70	0.40
Car pass.	0	0	0.30	0.20
Public	0.02	0	0	0.40

Decision tree example

Continuous choice



Choice of attribute variables

- Household/individual/situational attributes
- Attributes of evolving schedule
- Attributes of evolving schedule of partner
- Space-time opportunities (accessibility, time windows)
- Attributes of choice alternatives
- Availability of choice alternatives (constraints)

Making decisions (1)

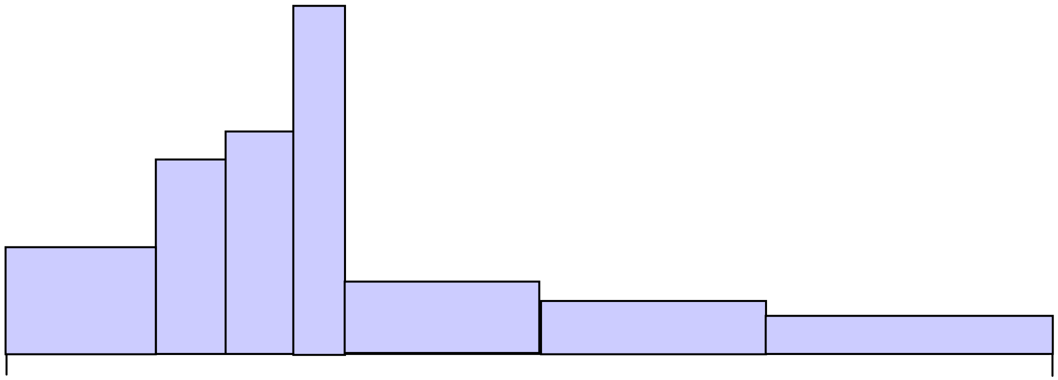
- Discrete choice

$$p_{ij} = \delta_{ij} \left(\frac{q_i}{\sum_i \delta_{ij} q_i} \right)$$

p_{ij}	probability of predicting choice *i* in case *j* (leaf node *k*)
q_i	probability of choice *i* in training set (leaf node *k*)
δ_{ij}	zero/one availability of choice *i* in case *j*

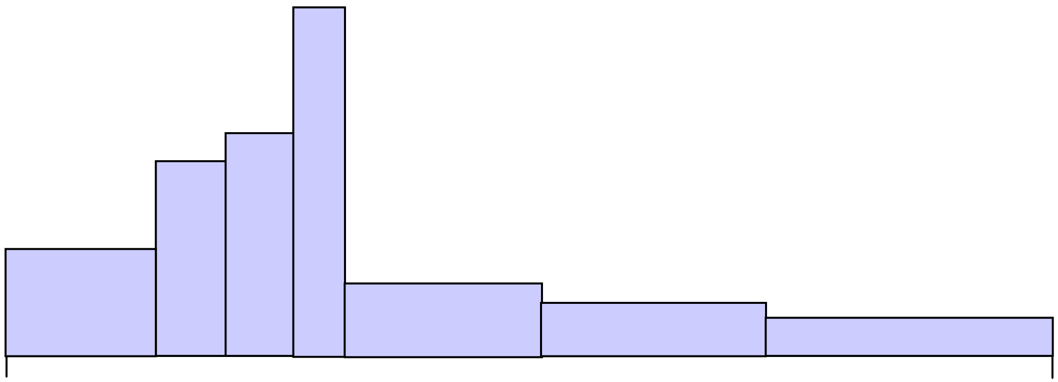
Making decisions (2)

- Continuous choice
  - Distributions often deviate strongly from normal form
  - Therefore we use equal frequency intervals to describe distributions



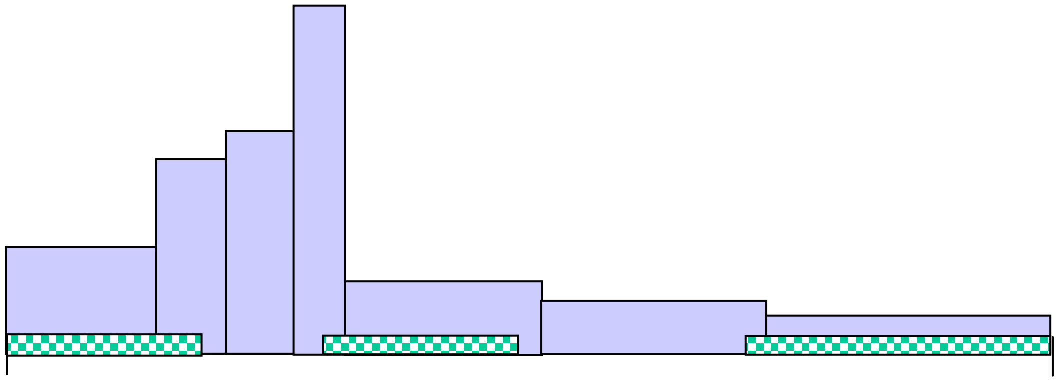
$$P_j(y) = \sum_i \Pr(i)\Pr(y \mid i)$$

$P_j(y)$	probability of drawing y in case j
$\Pr(i)$	probability of drawing EFI i
$\Pr(y \mid i)$	probability of drawing y given EFI i



$$\Pr(i) = \frac{1}{m} \qquad \Pr(y \mid i) = \sum_i \frac{\delta_i(y)}{d_i}$$

m	is number of EFI's
$\delta_i(y)$	0/1 y falls in EFI i
d_i	width of EFI i



$$P_j(y) = \frac{1}{m - \sum_i \left(\frac{b_{ij}}{d_i} \right)} \sum_i \left\{ \frac{\delta_{ij}(y)}{d_i} \right\}$$

Why decision tree induction ?

- Is able to represent a wider range of decision rules than just utility-maximization
- No pre-selection of attribute variables
- Completeness and consistency of rule set is guaranteed
- No assumptions regarding model form and distributions of variables
- Non-systematic variance can be reproduced in predictions

Space-time constraints

- Maximally available time window is defined based on:
  - Minimum activity duration
  - Nearest location of facilities
  - Fastest available transport mode
  - Opening hours of facilities
- The time window for schedule position i is found by shifting $j < i$ as far as possible to the left on the time scale and $j > i$ as far as possible to the right

Some key space-time condition variables (1)

Municipality choice

Orde0	Order of home municipality
Orde1	Order of chosen municipality
Maxb *j*	Highest order available in band *j*
Dbz *i*	Nearest band for order *i*
TRpuca *i*	Travel time ratio public-transport/car to nearest municipality of order *i*
CRpuca *i*	Travel costs ratio public-transport/car to nearest municipality of order *i*
Cpark *i*	Mean parking tariff in nearest municipality of order *i*
Avb *j*	Availability of chosen order in band *j*
Availo *i*	Accessibility of nearest order *i* given time-window for the activity
Availb *j*	Accessibility of band *j* given time-window for the activity
Avgem	Availability of a non-home municipality given time-window for the activity

Some key space-time condition variables (2)

Zone choice	
Zorde0	Order of home zone
Gorde0	Order of home municipality
Avo *i*	Availability of order *i* in municipality
Ddbz *i*	Car distance to nearest zone of order *I*
Avord *i*	Availability of order *i* given time window and choice of municipality
Avad *j*	Availability of chosen order in band *j*
Avzon *j*	Availability of zone *j* given time window and choice of municipality
TRpuca *i*	Travel time ratio public-transport/car to nearest municipality of order *i*
CRpuca *i*	Travel costs ratio public-transport/car to nearest municipality of order *i*
TRvona	Access/egress time as a ratio of total public transport
Cpark *i*	Mean parking tariff in nearest municipality of order *I*

Some key space-time condition variables (3)

Transport mode choice	
Dcar	Car distance
CRpuca	Travel cost ratio car / public-transport
TRpubi	Travel time ratio slow / public transport
TRpuca	Travel time ratio public transport / car
TRcoff	Travel time ratio congested / free floating condition
TRvona	Ratio of access and egress time of total public transport travel time
PRbeta	Access/egress time as a ratio of total public transport
Cpark	Mean parking tariff of paid parking places
trcon	There is a train connection

Study area data (1)

- Zoning systems
  - 4 Digit zip code areas (*n* = 3,987)
  - Municipalities (*n* = 625)
  - LMS subzones (*n* = 1308)
  - LMS zones (*n* = 345)
- Employment by sector (total, schools, services, daily, non-daily, leisure)
- Population (social activities)

Study area data (2)

- Transport system
  - Road network (type, distance, speed by mode)
  - Congested travel times
  - Bus/tram/metro (tariff zones, travel time, access + egress time, distance)
  - Train (travel time, access+egress time, distance)
- Car parking
  - Capacity free
  - Capacity paid
  - Mean price

Study area data (3)

- Opening times
    - Modal/largest opening and closing hours by sector

Activity diary data

- Four activity data sets from 4 surveys conducted in the NL were pooled (1997 – 2001)

  - 2 days activity diaries
  - Pre-coded scheme for activity reporting
  - Balanced across days of the week

- The pooled data set includes 6748 household-days and 9985 person-days

Decision tree induction results

- Together the 27 decision trees describe 1687 conditional choice probability distributions

- Goodness-of-fit of the model was measured at:

  - Decision tree level
  - Activity pattern level (SAM)
  - Aggregate level

Goodness-of-fit decision trees (discrete)

DT id	DT label	nmin	ncond	nalt	nobs	nleaf	e_0	e	e_{incr}	c
3	Work/school	84	21	2	8455	49	0.506	0.772	0.538	0.593
5	# of work episodes	37	22	2	3757	20	0.640	0.670	0.083	0.303
9	fixed activity	90	38	2	35008	114	0.797	0.829	0.157	0.373
10	# of fix. activity epsiodes	40	38	4	4003	24	0.471	0.524	0.101	0.196
12	Fix act. on work trip	30	38	5	2656	39	0.422	0.488	0.114	0.578
14	*L* same as previous	55	33	2	5579	54	0.518	0.625	0.222	0.432
15	*L* municipality. in/out	90	29	2	18758	105	0.512	0.625	0.277	0.468
16	*L* municipality order	79	43	5	7932	63	0.229	0.304	0.097	0.525
17	*L* municipality nearest	79	38	2	7932	55	0.503	0.727	0.451	0.560
18	*L* muninicipality distance band	42	43	6	4279	67	0.168	0.331	0.196	0.715
19	*L* zone order	90	40	4	17782	127	0.260	0.385	0.169	0.577
20	*L* zone distance band	90	47	5	9510	68	0.258	0.422	0.221	0.672
21	Mode to work	36	39	4	3665	51	0.381	0.590	0.338	0.659
22	flexible activity	90	49	2	62164	204	0.672	0.734	0.190	0.405
23	With whom flex. act.	90	49	3	12899	86	0.364	0.500	0.214	0.552
24	Duration flex. act.	90	51	3	12899	71	0.342	0.389	0.071	0.356
25	Start time flex. Act.	90	63	6	12709	87	0.174	0.335	0.195	0.693
26	Trip chaining	90	48	4	11107	46	0.484	0.785	0.584	0.801
27	Mode to non-work	90	38	4	9523	56	0.425	0.607	0.317	0.614

$$e = \frac{1}{n}\sum_k \frac{\sum_i (f_{ik})^2}{f_k}$$

Goodness-of-fit decision trees (continuous)

DT id	DT label	nmin	ncond	nobs	nleaf	s_0'	s'	s_{incr}	F
1	End time Sleep	55	20	8372	51	8.54 e-03	5.26 e-03	0.0346	76.07
2	Start time sleep	55	20	8372	62	7.40 e-03	3.85 e-03	0.0385	133.75
4	Duration work	30	21	3757	37	3.60 e-03	2.95 e-03	0.0067	26.36
6	Duration ratio work episodes	30	22	900	1	1.52 e-02	1.52 e-02	0	0
7	Duration work break	30	23	900	8	1.31 e-02	8.62 e-03	0.0495	36.94
8	Start time work	30	39	3757	41	8.44 e-03	5.70 e-03	0.0291	17.57
11	Duration fixed act.	34	38	5120	43	2.06 e-03	1.04 e-03	0.0102	37.30
13	Start time fixed act.	40	38	6065	58	2.89 e-02	2.41 e-02	0.0633	73.44

$$s' = \frac{1}{m^2}\frac{1}{n}\sum_k f_k \sum_i \frac{1}{d_{ik}+1}$$

Goodness-of-fit pattern level

	Mean	St.dev.
SAM atype	5.194	3.035
SAM with	6.714	3.479
SAM loc	3.467	2.486
SAM mode	5.906	3.659
UDSAM	26.475	13.642
MDSAM	12.205	6.417

SAM minimum effort required to make observed and predicted pattern identical by insertion, deletion and substitution operations

Goodness-of-fit aggregate level (1)

	Df	Rel. diff.	c
# of work activities	5	0.008	0.1014
# of sec. fixed activities	5	0.017	0.0805
# of flexible activities	5	0.025	0.0728
Total # of sec. activities	5	0.027	0.0755
# of tours	5	0.050	0.1386
# of activities in tour	4	0.025	0.0815
Mode of first link	3	0.041	0.0659
Activity type	9	0.020	0.0933
Mode	3	0.064	0.1054

Goodness-of-fit aggregate level (2)

	Df	Rel. diff.	c
Time of day	5	0.015	0.0414
Duration (flex.)	2	0.042	0.0425
Travel party (flex.)	2	0.033	0.0350
Trip chaining	3	0.042	0.0716
Municipality		0.018	0.0089
Mun. order (extern)	4	0.017	0.0407
Distance (extern)	9	0.020	0.1132
Distance (extern)	9	0.007	0.0456
Mun. population (extern)	9	0.018	0.0887
Zone employment	9	0.012	0.0621

Goodness-of-fit aggregate level (3)

	Δ m/m0	t-value
Work duration	-0.0048	-0.757
Distance (extern)	-0.0821	-5.233
Distance (intern)	0.0963	5.313
Mun. population (extern)	-0.0778	-4.226
Zone employment	-0.0972	-8.796

Some conclusions (1)

- Predictability of decisions varies strongly across choice facets

- Relatively well predictable are:

  - Y/n work activity
  - Y/n secondary fixed activity
  - Relative location
  - Nearest/other municipality
  - Y/n flexible activity

Some conclusions (2)

- Poorly predictable are
  - Municipality order
  - Municipality distance band
  - Zone order
  - Flexible activity duration
  - Flexible activity start time
- *Relative* performance is high for
  - Y/n work activity
  - Nearest/other municipality
  - Trip chaining

Some conclusions (3)

- Generally, predictions at aggregate level are unbiased
- In particular, the location module performs very satisfactory
- Exceptions
  - Overprediction of number of flexible activities and number of tours
  - Slight underprediction of slow mode
  - Slight overprediction of activities after 6 PM

Example of a decision tree (Mode choice, Part 1)

Urb	-	-	-	-	-	-	-	-	-	-	0,2,3,4	1	-	-	-	-	-	-	-	-
Comp	-	-	-	-	0,3,1	2,4	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SEC	-	0,3,1	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ncar	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	2	2	2	0	0
Gend	-	-	-	-	0	0	0	0	0	0	0	0	1	1	1	0	0	1	-	-
Driver	0	1	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	1
wstat	-	0,1	0,1	2	-	-	-	-	-	-	-	-	-	0,1	2	-	-	-	-	-
Pwstat	-	-	-	-	0,1	0,1	0,1	2	2	2	-	-	-	-	-	-	-	-	-	-
Nsec	-	-	-	-	0-3	0-3	5-4	-	0-3	5-4	-	-	-	-	-	-	-	-	-	-
Adur1	-	-	-	-	0	0	0	0	0	0	1,2	1,2	-	-	-	0	1,2	-	-	-
Cbrget	-	-	-	-	-	-	-	0	1	1	-	-	-	-	-	-	-	-	-	-
Dist	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Pstat	-	-	-	-	-	-	-	-	-	-	-	-	0	1,3,2	1,3,2	-	-	-	-	-
slow	0.99	0.95	0.87	0.78	0.85	0.95	0.70	0.94	0.87	0.96	0.85	0.71	0.72	0.87	0.79	0.80	0.65	0.55	0.97	0.76
car	0.00	0.03	0.11	0.21	0.14	0.02	0.29	0.05	0.13	0.04	0.12	0.16	0.27	0.13	0.19	0.16	0.29	0.42	0.00	0.16
pub	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.04
pass	0.01	0.03	0.02	0.01	0.01	0.02	0.01	0.01	0.00	0.00	0.04	0.12	0.00	0.00	0.02	0.03	0.05	0.04	0.02	0.04
N	255	222	109	97	144	149	142	150	157	235	523	161	235	260	295	250	170	227	108	198
	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20

Interpreting decision trees: Impact tables

- Decision trees derived from data are generally complex and difficult to interpret

- In post-processing stage, the impact of each condition variable on choice is measured based on a sensitivity analysis

$$IS_s = D\left(\boldsymbol{F}_s, \overline{\boldsymbol{F}_s}\right)$$

Impact of *s* on choice distribution

$$IS_{si} = D\left(\boldsymbol{F}_{si}, \overline{\boldsymbol{F}_{si}}\right)$$

Impact of *s* on choice *i*

$$MS_{si} = \frac{\sum_{u(s)=2}\left(f_{u(s),i} - f_{u(s)-1,i}\right)}{\sum_{u(s)=2}\left|f_{u(s),i} - f_{u(s)-1,i}\right|}$$

Monotonicity of impact of *s* on choice *i*

Example of an impact table (mode choice)

Cond	*IS*	*IS*slow	*IS*car	*IS*pub	*IS*pass	*MS*slow	*MS*car	*MS*pub	*MS*pass
Urb	7.13	1.04	0.76	1.51	3.83	-0.13	0.44	-0.86	-0.06
Comp	0.69	0.23	0.43	0.03	0.00	0.33	-0.33	0.33	0.33
Age	6.80	0.22	1.61	0.03	4.94	0.20	-1.00	1.00	1.00
Ncar	2158.20	845.92	1146.02	110.51	55.74	-1.00	1.00	-0.97	1.00
Gend	426.32	10.34	148.09	3.56	264.34	-1.00	1.00	-1.00	-1.00
Driver	7.16	3.00	3.82	0.08	0.26	-1.00	1.00	-1.00	1.00
wstat	11.35	3.91	6.95	0.01	0.48	-1.00	1.00	1.00	-0.78
Pwstat	3.33	1.46	1.81	0.01	0.05	1.00	-1.00	-1.00	-1.00
Nsec	1.75	0.59	1.12	0.02	0.03	-1.00	1.00	-1.00	-1.00
Adur1	50.58	17.17	6.23	3.31	23.87	-1.00	0.66	0.82	1.00
Cbrget	0.68	0.24	0.09	0.00	0.35	1.00	-1.00		-1.00
Cadist	25167.09	15507.72	3608.72	2858.38	3192.44	-1.00	0.82	1.00	0.88
TRvona	19.31	8.76	6.69	0.18	3.68	1.00	-1.00	-1.00	-1.00
PRbeta	285.48	23.30	46.73	182.44	33.01	1.00	-1.00	1.00	-1.00
Cpark	136.97	51.28	60.26	7.13	18.29	1.00	-1.00	1.00	-1.00
Pstat	57.40	1.65	16.35	0.64	38.75	0.60	-0.16	-0.29	0.01
Pdist	12.04	0.08	0.67	1.99	9.29	-1.00	-1.00	-1.00	1.00

Conclusions

- The model consists of 27 linked decision trees, a total of 1687 conditional choice probability distributions derived from 9985 person-day diaries
- Activity patterns are predicted from scratch
- The model uses travel time, travel distance, travel costs, land-use and parking data for the whole of the Netherlands
- Predicted activity patterns should not violate space-time and situational constraints
- Residual variance is reproduced in predictions; aggregate distributions are almost bias free