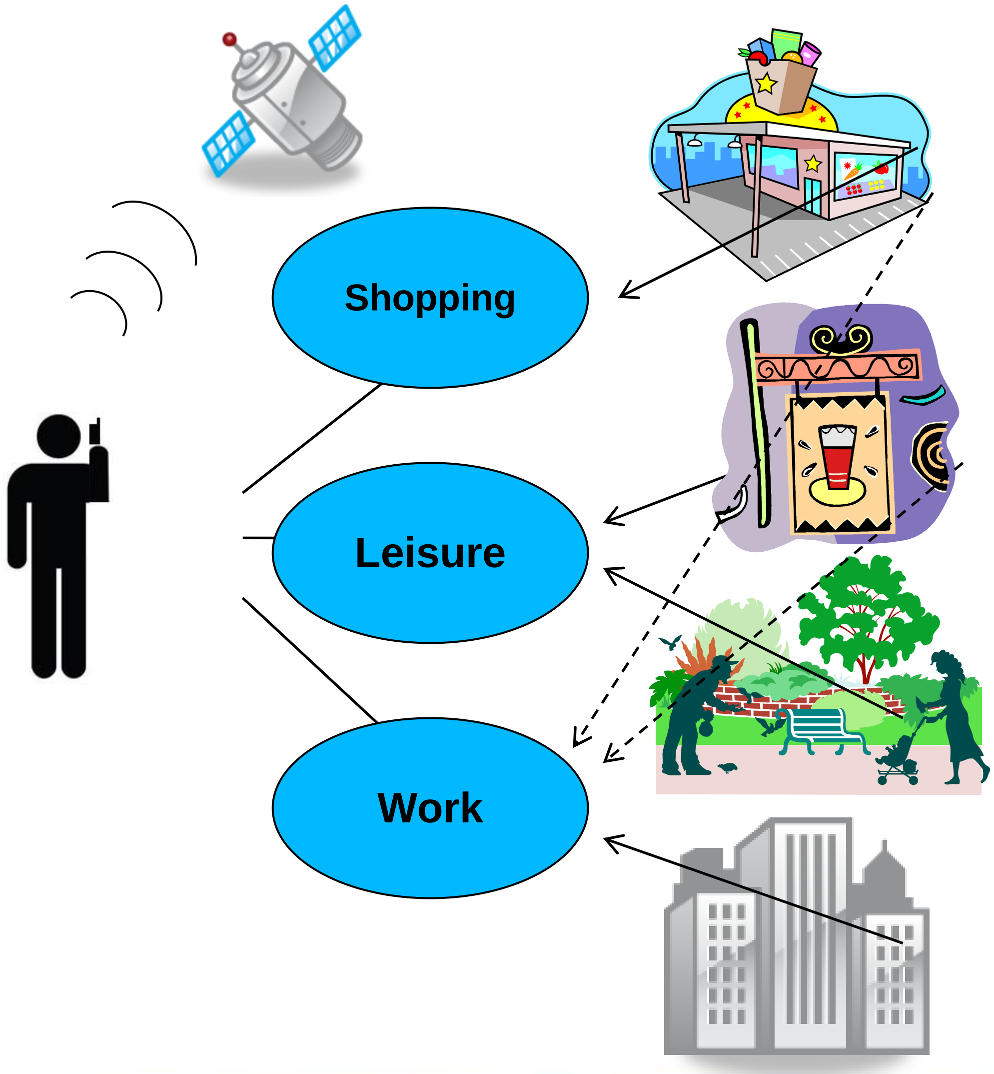
Inferring activity choice from context measurements using Bayesian inference and random utility models

Ricardo Hurtubia
Gunnar Flötteröd
Michel Bierlaire

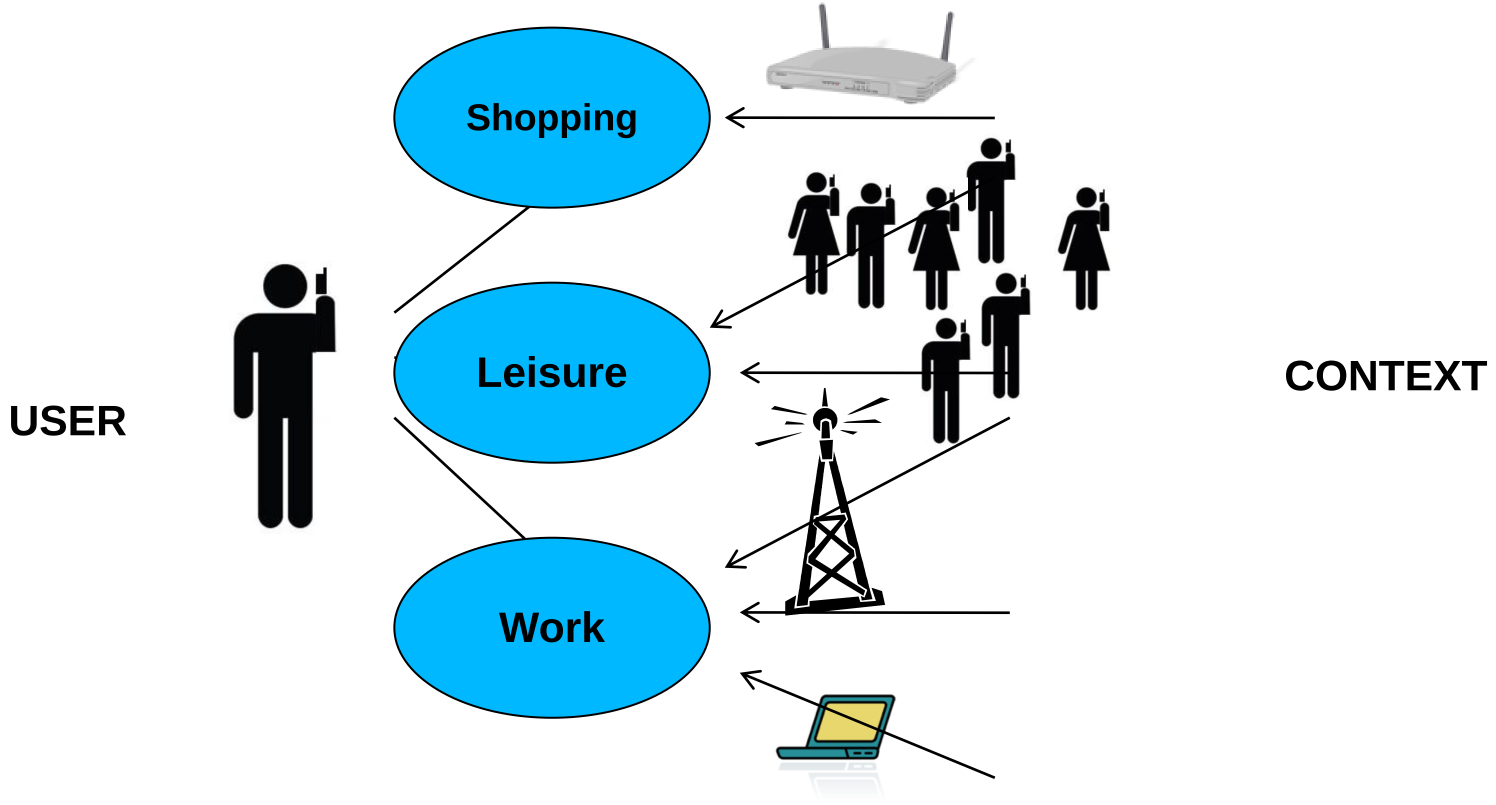
Transport and mobility laboratory - EPFL

Urbanics – February 2010

Motivation



Motivation



Outline

1. Motivation
2. Framework
   2.1 Prior model
   2.2 Collected data
   2.3 Likelihood function
3. Results / Case study
4. Practical Issues
5. Possible improvements
6. Conclusions

General framework

- Objective: combine general knowledge of population's behavior and individual context variables' measurements into estimates of an individual's activities
- Available data:
  - Reported activities in Swiss Transport Microcensus 2005
  - Land use data
  - Measurements from a smartphone for one user over a two-month period
  - Activity survey
- Bayesian inference:

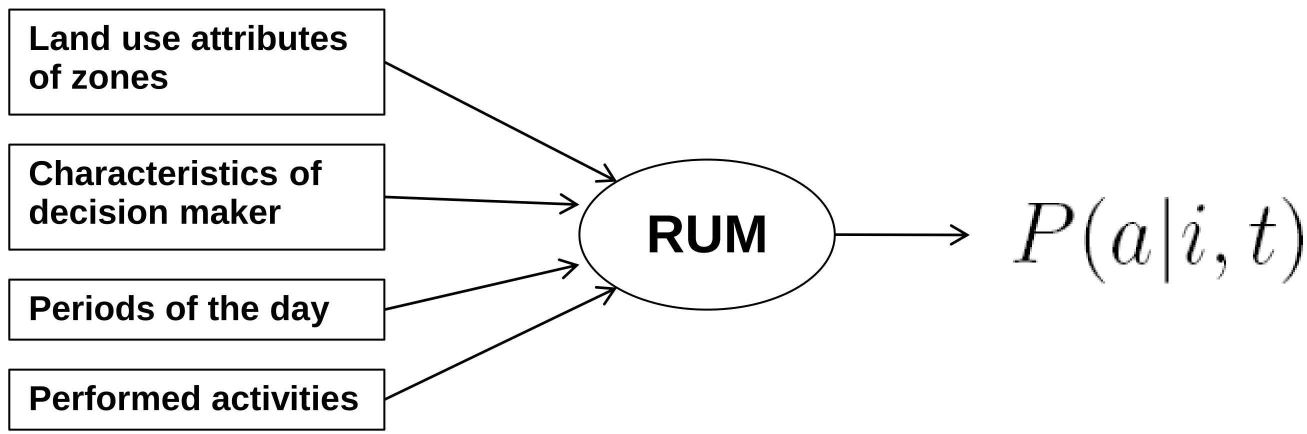
$$P(activity|measurements) \propto P(activity) \cdot P(measurements|activity)$$

Prior: $P(activity)$

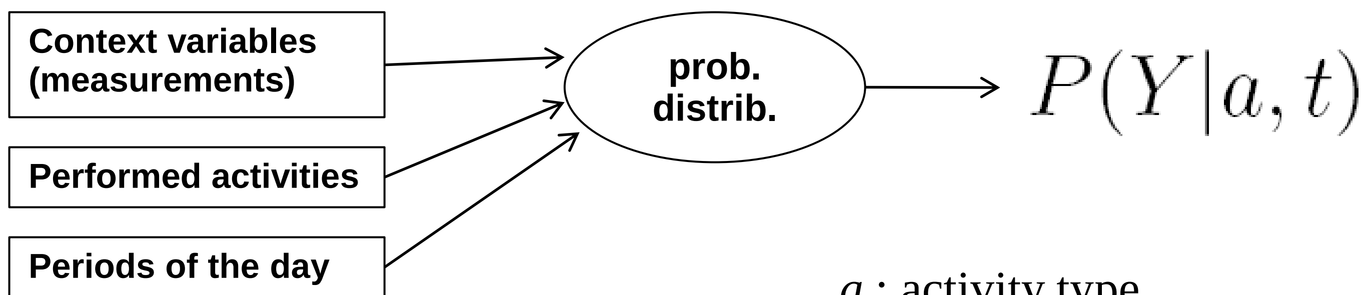
Likelihood: $P(measurements|activity)$

General framework

- Prior:



- Likelihood:



a : activity type
i : zone
t : period
Y: measurement

Prior model

- Probability of performing a certain type of activity given a location (zone) and a time of the day
- Structure: Multinomial logit

$$P_n(a \mid i,t) = \frac{\exp(U_{na}(z_i, z_n, \delta_t))}{\sum_{a'} \exp(U_{na'}(z_i, z_n, \delta_t))}$$

a : type of activity (work, study, leisure, shopping….)
z_i : land use attributes of zone i
z_n : attributes of user n
δ_t : indicator of the period of the day {morning, noon, afternoon, night}

Prior model estimation results

	parameter	work	study	shopping	services	leisure	other
	constant	-	**-0.532**	**2.031**	**2.311**	**3.522**	**0.656**
n	male	**0.713**	-	**-0.377**	**-0.278**	-	-
n	employed	**2.132**	-	-	-	-	-
n	children	-	-	-	-	-	**0.379***
t	morning	**2.720**	-	**0.887**	**1.341**	-	-
t	noon	**1.001**	-	-	-	-	-
i	industry	**0.025**	-	-	-	-	-
i	commerce	-	-	**0.077**	-	-	-
i	services	**0.046**	-	-	**0.055**	**0.024**	-
i	other	**0.032**	-	-	-	**0.053**	**0.065***
i	retail	-	-	**1.074**	-	-	-
i	long term retail	-	-	**0.554**	-	-	-
i	restaurant	-	-	-	-	**0.109**	-
i x n	school*age<19	-	**1.694**	-	-	-	-
i x n	high_educ*student	-	**1.328**	-	-	-	-
t x n	morning*student	-	**6.516**	-	-	-	-
t x n	noon*student	-	**4.212**	-	-	-	-
t x n	morning*age>60	-	-	**1.114**	-	**0.836**	-
t x n	afternoon*age<19	-	-	-	-	**0.813**	-
t x n	afternoon*age>60	-	-	-	-	**-0.242**	-
t x n	night*age19_25	-	-	-	-	**1.683**	-

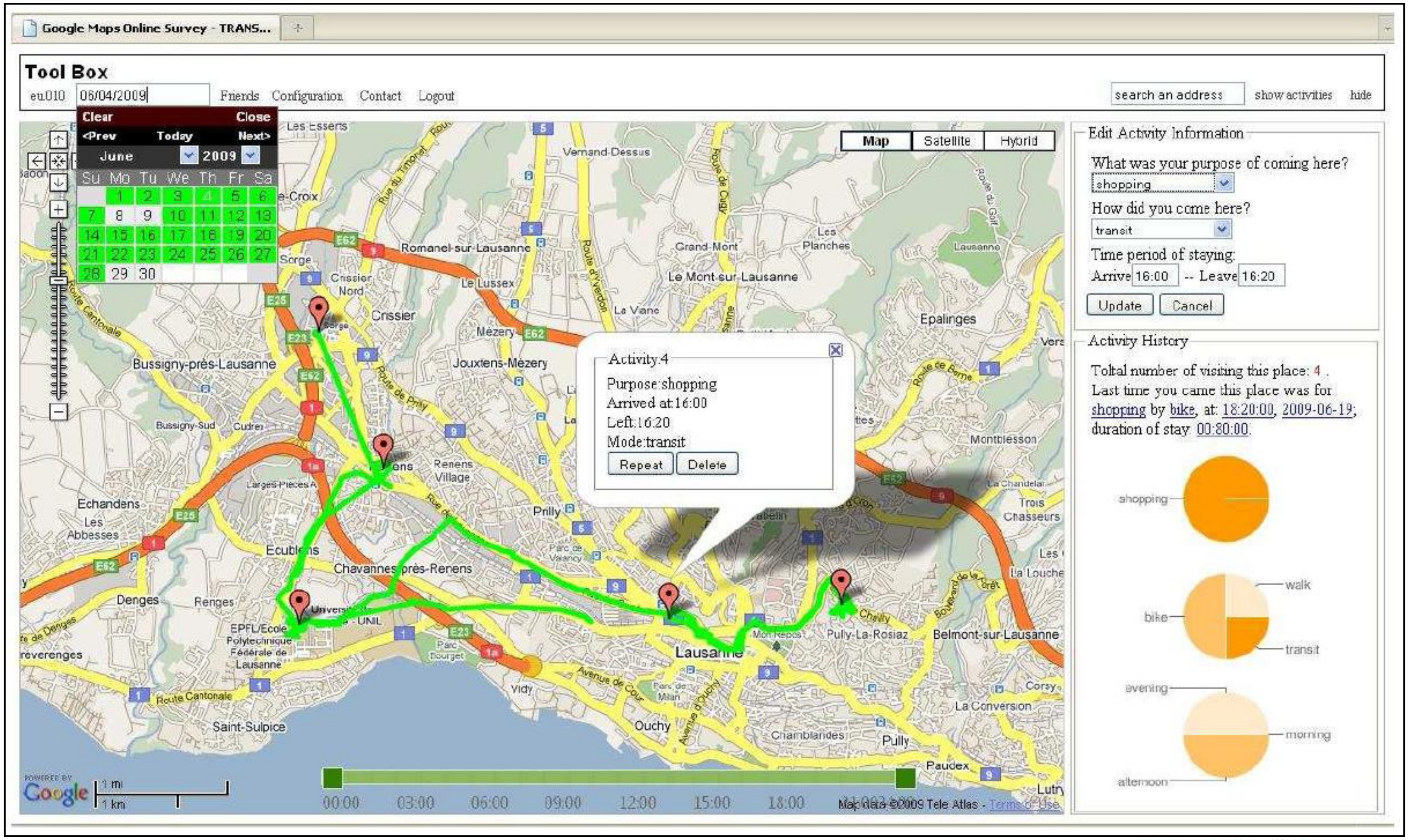
estimated using Biogeme (Bierlaire, 2003)

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

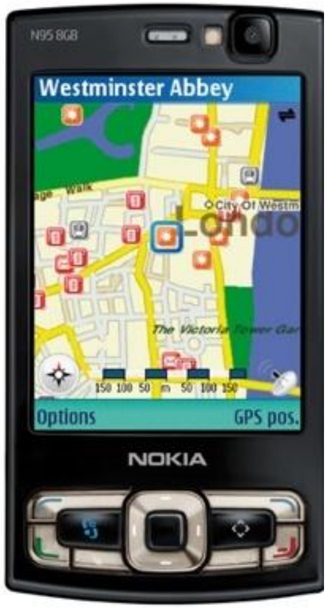
Data: survey

- Daily activity survey: Two months, one user
  - Location
  - Time
  - Type of activity
  - Transport mode



Data: measurements

- Measurements from a smartphone (Nokia N95)
- Context variables:
  - GPS location
  - Nearby networks (LAN,GPRS, cell id)
  - Nearby Bluetooth devices (MAC address)
  - Movement detection (accelerometer)
  - Call log (duration, direction, contact)
  - SMS (length, direction, contact)
  - Camera usage
  - Media player usage
  - Profile (silent, general, etc)
  - Battery life
  - Energy plug state
  - Inactive time
  - …….



Data: measurements

Types of measurements:

- Active: actions triggered by the user
  - Camera
  - Media player
  - Calls/SMS
  - Profile

- Passive: product of the environment
  - Detected networks
  - Nearby bluetooth devices

TRANSP-OR

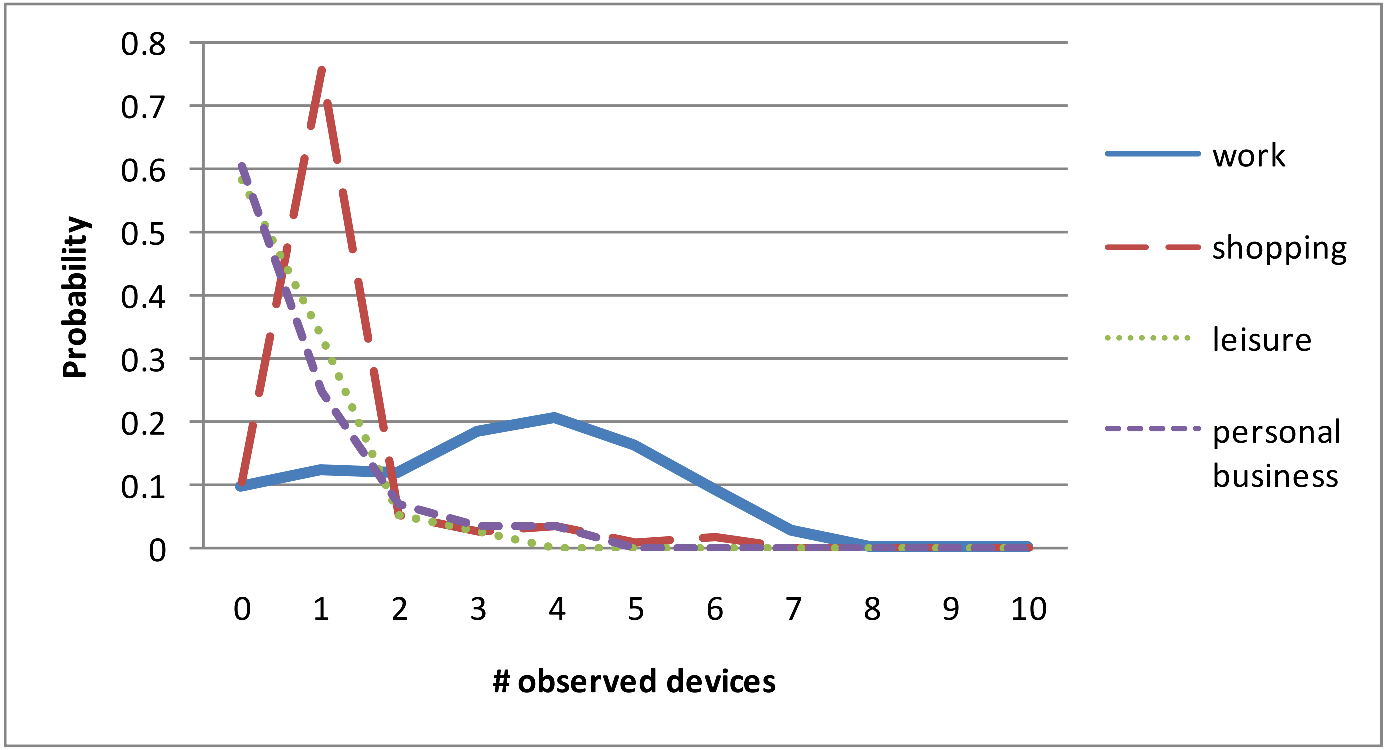
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Data: measurements

- Bluetooth measurements can be understood as either an active or passive measurement
  - Number of nearby Bluetooth devices → passive
  - Detection of particular devices → active
    - Decision of user to perform certain activity with specific individuals
    - Decision of other individuals to perform activities with user

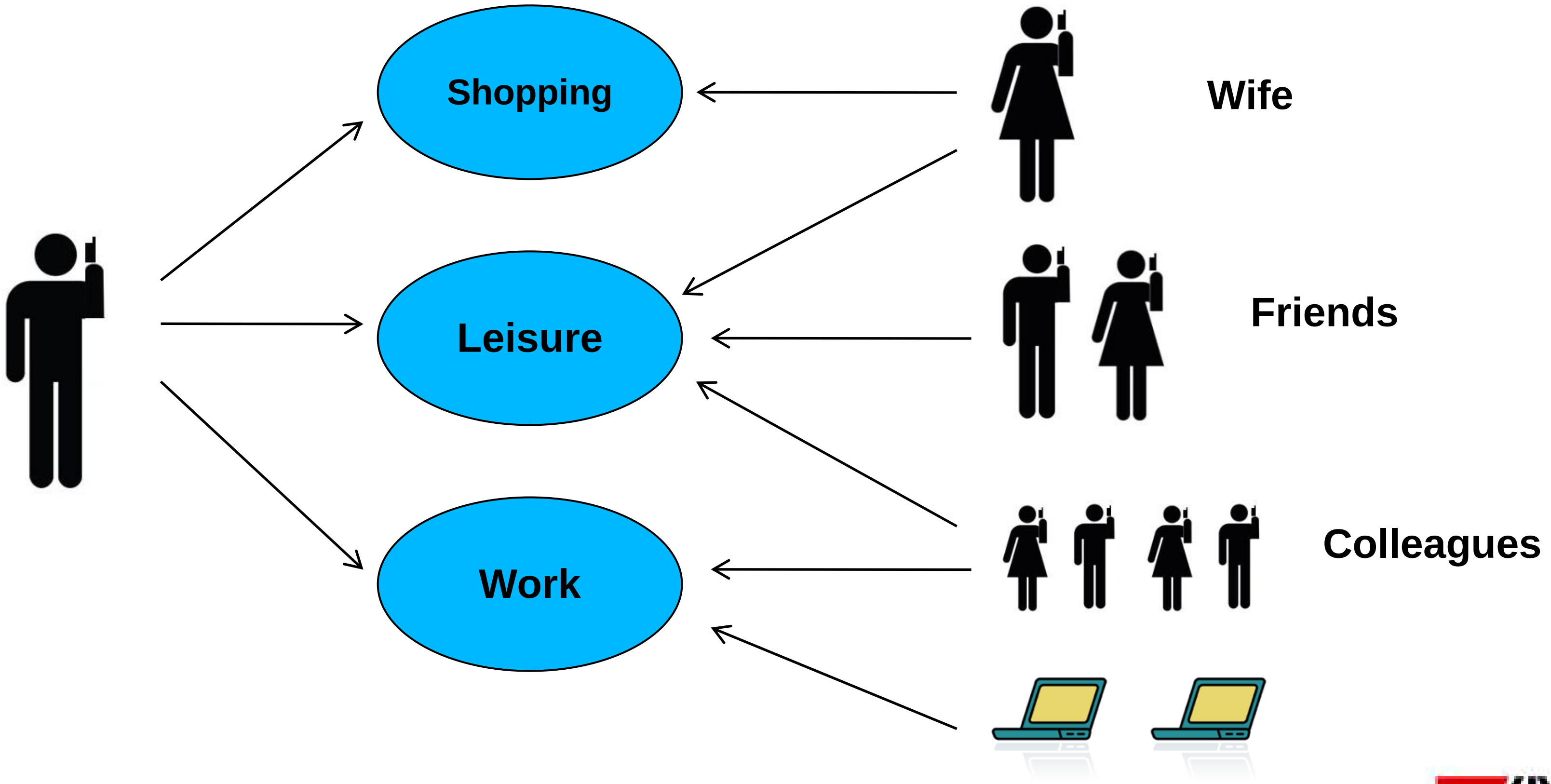
Measurements: Bluetooth (passive)

- Aprox 8700 measurements
- Distribution of number of detected devices:



Measurements: Bluetooth (active)

Frequent Bluetooth devices: some devices are mostly observed when performing certain types of activities



Likelihood

- We define

$$y_j = \begin{cases} 1 & \textbf{if device j is observed} \\ 0 & \end{cases}$$

- Joint likelihood:

$$P(Y|a,t) = \prod_j (P(y_j = 1|a,t) \cdot y_j + (1 - P(y_j = 1|a,t)) \cdot (1 - y_j))$$

Probability of observing device *j*

Probability of not observing device *j*

Likelihood

- Empirical probability of observing a device given the activity type and period of the day:

$$P(y_j = 1 | a, t) = \frac{N_{jat} + \varepsilon_a \cdot \alpha}{N_{at} + \alpha}$$

where:

– N_{at}: number of times activities type *a* were performed during period *t*

– N_{jat} : number of times device *j* was detected while performing activities type *a* during *t*

– ε_a : probability of observing any device while performing activity type *a*

– α : weight of “uninformed prior knowledge”

Inference

- We update the prior using the likelihood of the Bluetooth devices' measurements

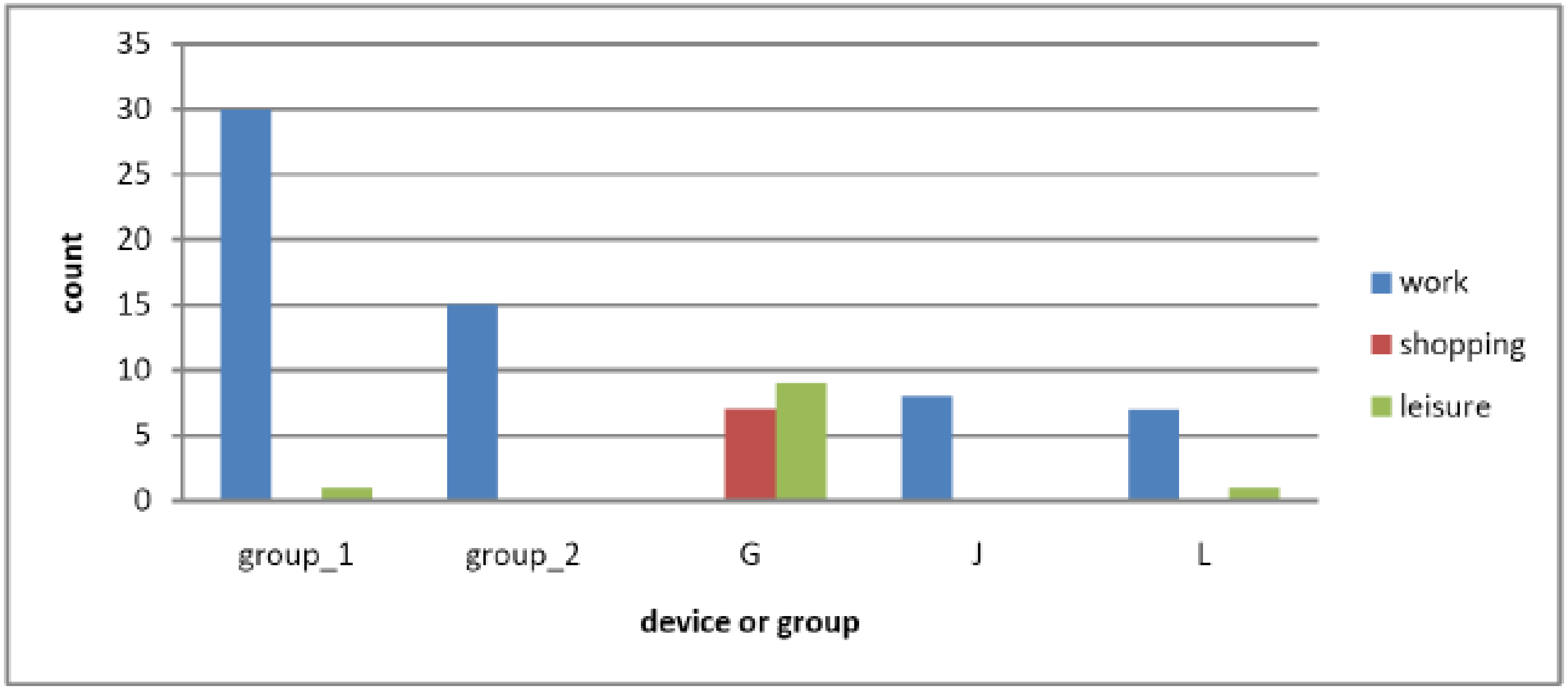
$$P(a|Y,i,t) = \frac{P(Y|a,t) \cdot P(a|i,t)}{P(Y|i,t)}$$

where:

$$P(Y|i,t) = \sum_{a'} P(Y|a',t) \cdot P(a'|i,t)$$

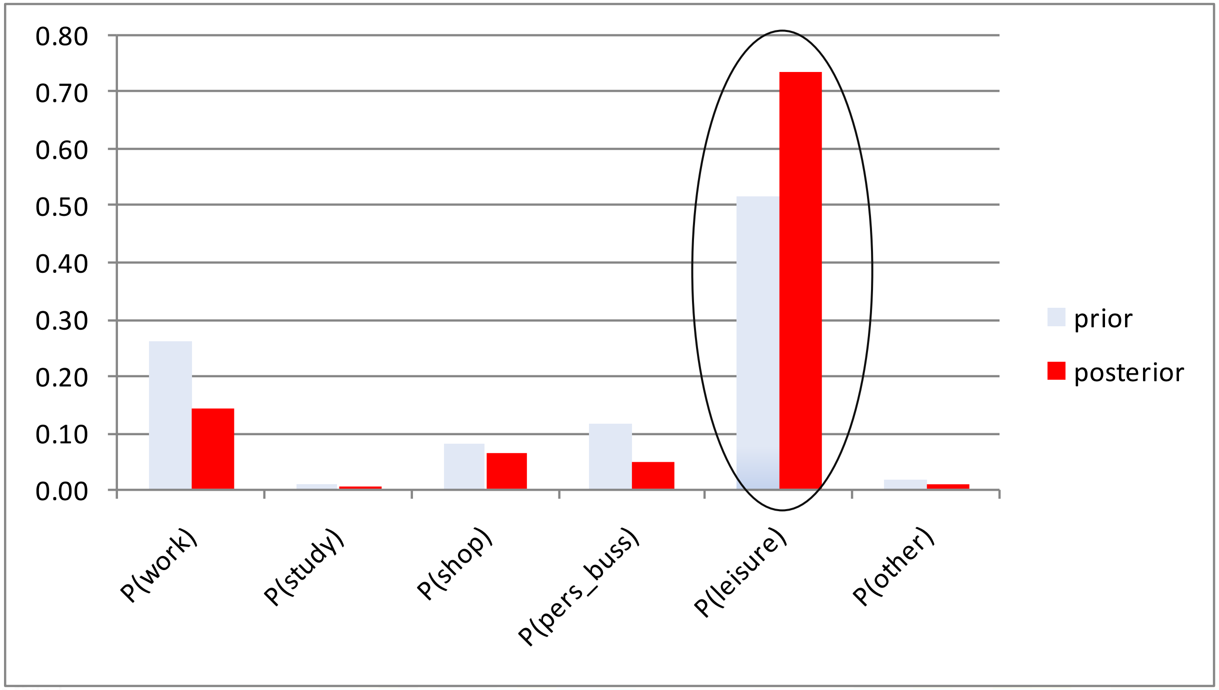
Case study

- 12 independent devices appear more than 4 times
- Grouped according to simultaneous-detection correlation



Case study

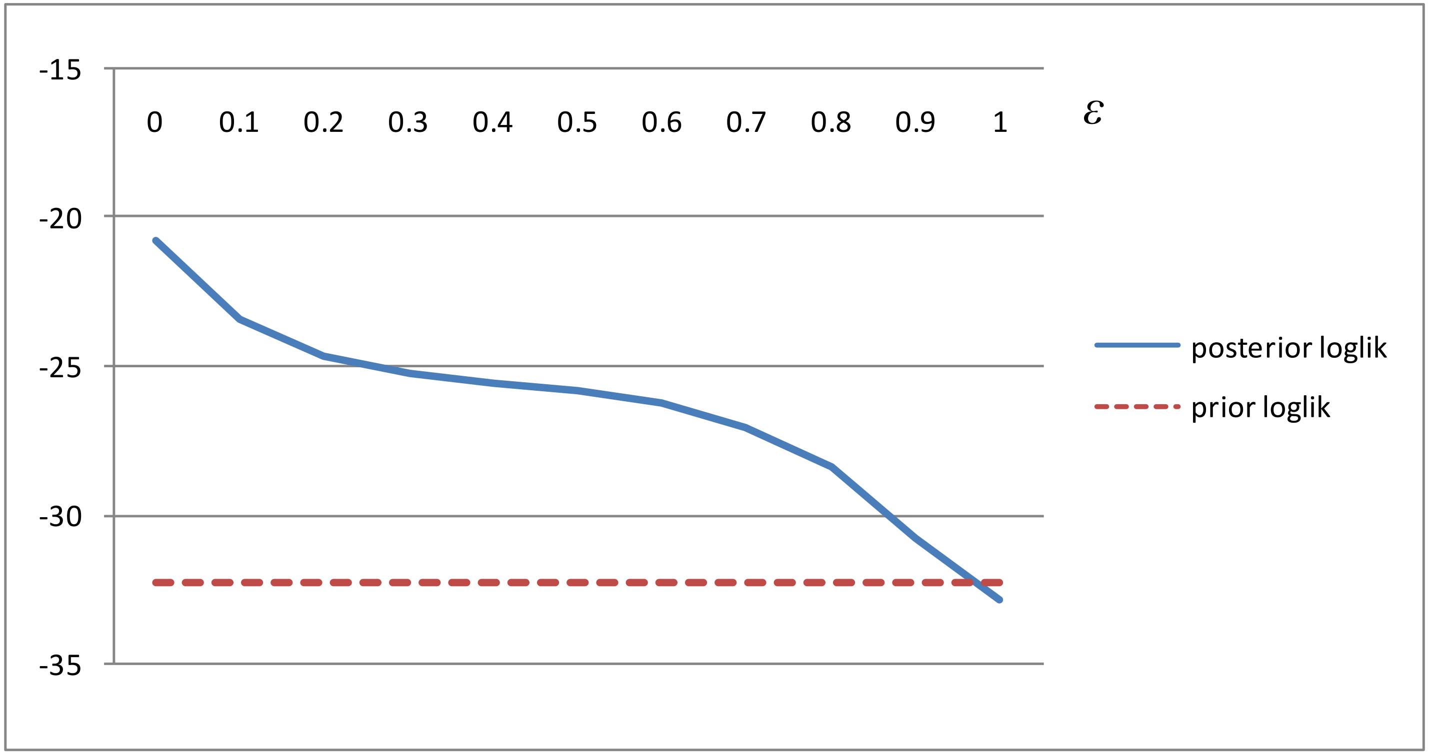
- A particular event
  - Leisure activity performed at work location during afternoon/night
  - Detection of devices:
    - Group_1 (frequent at work, also observed at leisure)
    - Device G (frequent at shopping and leisure, never observed at work)
    - Device J (observed only at work)



$\varepsilon = 0.01$
$\alpha = 10$

Case study

- If we assume a high value for epsilon, the aggregate fit of the posterior distribution deteriorates



$$\sum_{k}^{K} \log\left(\sum_{a} P_k(a) \cdot 1_{ak} \right)$$

Practical issues

- How do we update the likelihood if there is no survey information available?
  - "end of day" update: $P_k(a \mid i,t) \quad \forall k$

$$\left.\begin{aligned} N'_{jat} &= N_{jat} + \sum_k P_k(a \mid i,t) \cdot y_{jk} \\ N'_{at} &= N_{at} + \sum_k P_k(a \mid i,t) \end{aligned}\right\} \Rightarrow P'(y_j \mid a,t)$$

Practical issues

- Problem: endogeneity
  - Our prediction of the activity type depends on a likelihood function based on the same prediction
    - Potential propagation of errors
    - Generation of noise.
- Solution?
  - “pop-up questions”
    - Are you at work?
    - Are you shopping?
    - …

Possible improvements

- **Behavioral approach:**

  If we understand the measurement as user n choosing to perform activity *a* with user/device *j*:

$$P_n(y_j \mid a) = f(x_j, x_n, \beta_a)$$

  Possible attributes (x_j):
  - Previously observed frequency while performing *a*
  - Presence while performing other types of activity
  - Presence in different locations

- **Inclusion of other measurements in the likelihood function**

Conclusions and further work

- Bayesian approach allows to improve the quality of the activity type inference
- Bluetooth measurements are useful to infer activity type

Further work

- Test of “end of day update”
- Behavioral explanation of the likelihood function
- Inclusion of other measurements in the likelihood function

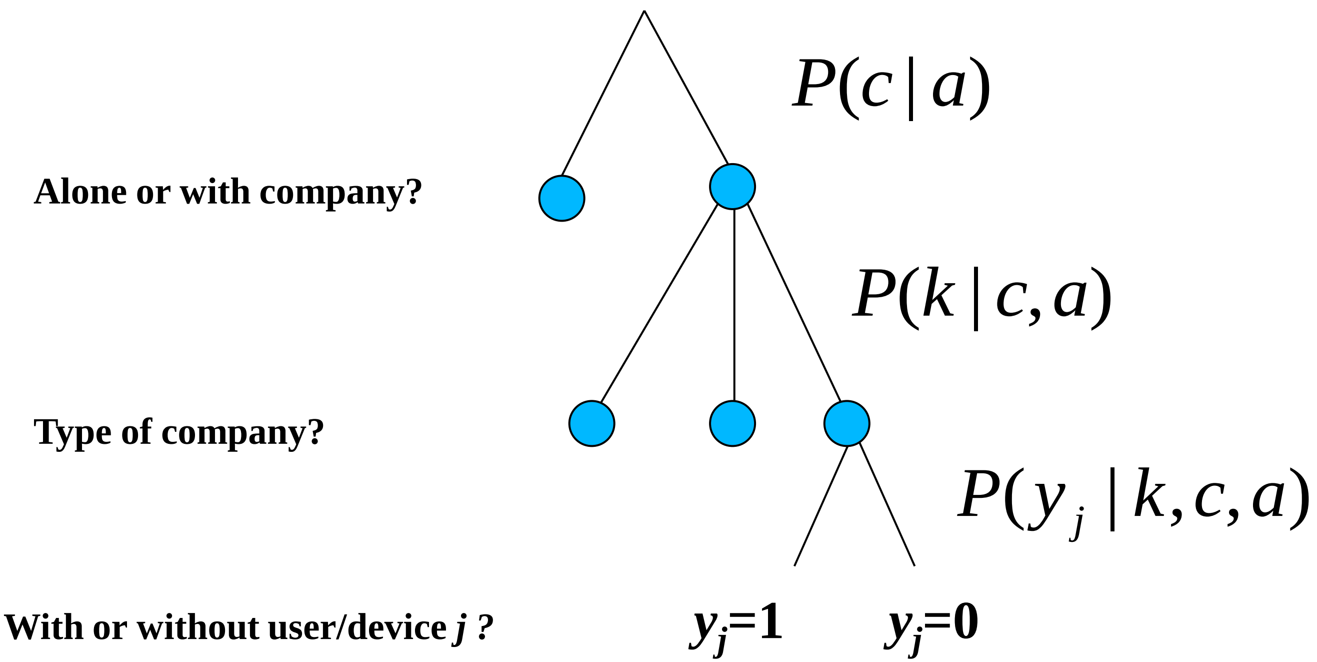
Thank you

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Behavioral approach

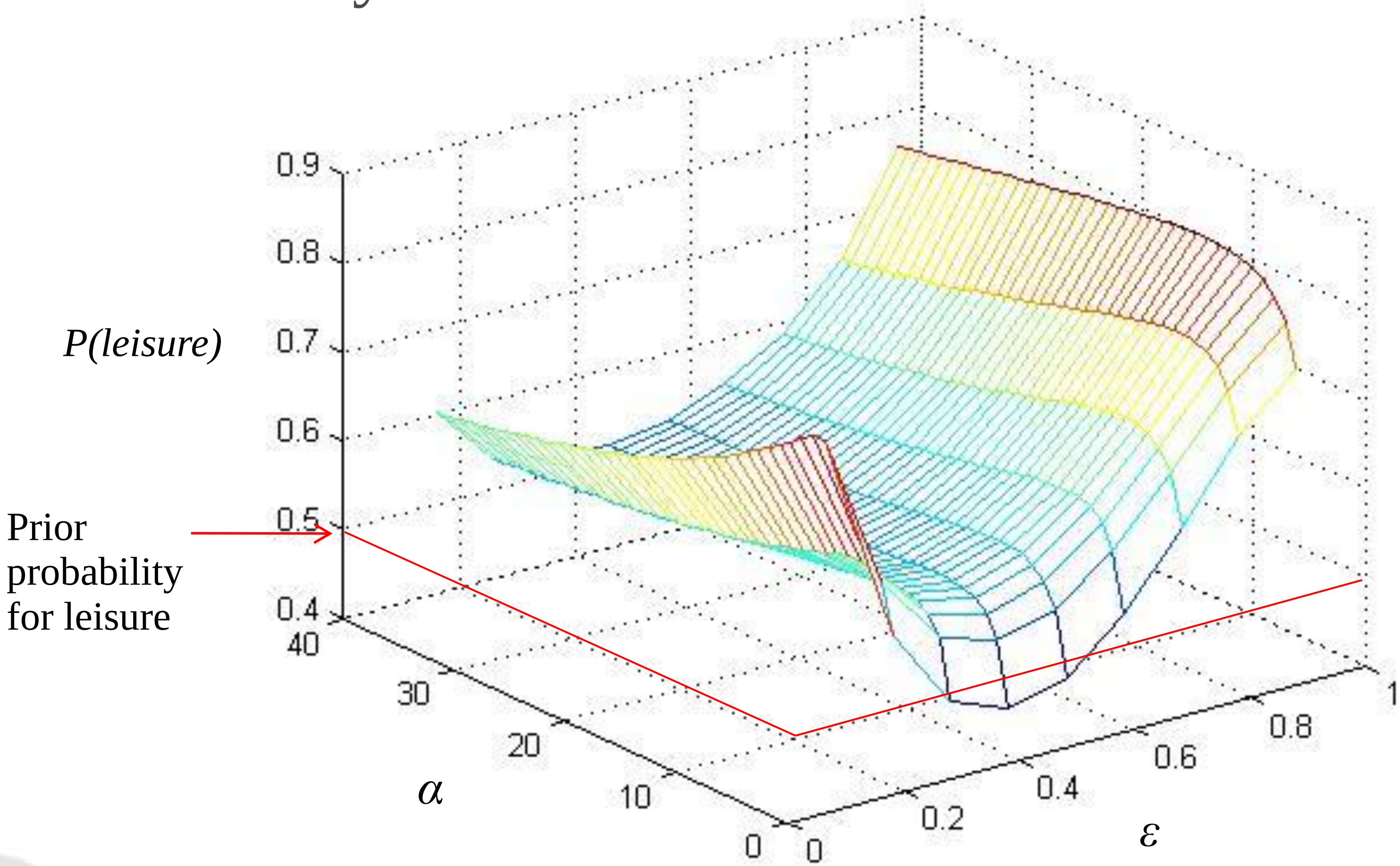
- A better behavioral explanation?



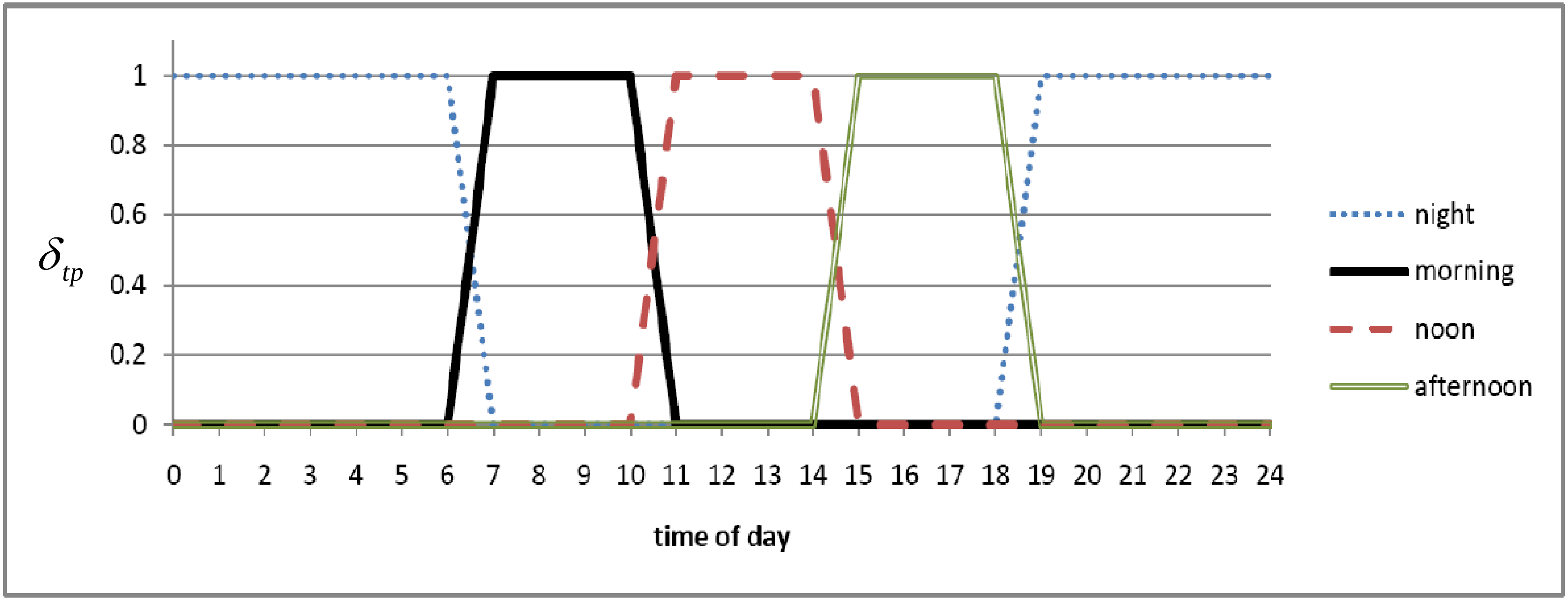
$$P(y_j|a) = \sum_c \sum_k P(y_j|k,c,a) \cdot P(k|c,a) \cdot P(c|a)$$

Case study

- Sensibility to α and ε.



Time discretization



$$\delta_t = (\delta_{tp}) \qquad p \in \{\text{night, morning, noon, afternoon}\}$$

Correlation of devices

correl	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	1	G1	G1	G1	G1	G1			G1					
B	0.73	1	G1	G1	G1	G1			G1					
C	0.79	0.78	1	G1	G1	G1			G1					
D	0.81	0.80	0.80	1	G1	G1			G1					
E	0.70	0.68	0.68	0.71	1	G1			G1					
F	0.73	0.59	0.65	0.79	0.60	1			G1					
G	-0.27	-0.25	-0.25	-0.25	-0.23	-0.23	1			G2				
H	0.51	0.61	0.48	0.57	0.40	0.49	-0.19	1				G3		
I	0.58	0.68	0.68	0.70	0.54	0.42	-0.19	0.13	1					
J	-0.26	-0.25	-0.25	-0.24	-0.22	-0.22	0.96	-0.18	-0.18	1				
K	0.41	0.52	0.52	0.54	0.48	0.40	-0.13	0.49	0.29	-0.13	1			
L	0.50	0.52	0.44	0.54	0.39	0.50	-0.13	0.70	0.08	-0.13	0.59	1		
M	0.41	0.44	0.35	0.45	0.30	0.31	-0.13	0.18	0.39	-0.13	0.32	0.18	1	
N	-0.50	-0.47	-0.47	-0.46	-0.43	-0.37	0.54	-0.35	-0.35	0.52	-0.25	-0.25	-0.17	1.00

$$correl(j, j^*) = \frac{\sum (y_j - \overline{y_j})(y_{j^*} - \overline{y_{j^*}})}{\sqrt{\sum (y_j - \overline{y_j})^2 \sum (y_{j^*} - \overline{y_{j^*}})^2}}$$

BACK