Three challenges in route choice modeling

Michel Bierlaire and Emma Frejinger

`transp-or.epfl.ch`

Transport and Mobility Laboratory, EPFL

Route choice modeling

Given a transportation network composed of nodes, links, origin and destinations.
For a given transportation mode and origin-destination pair, which is the chosen route?

Applications

- Intelligent transportation systems
- GPS navigation
- Transportation planning

Challenges

- Alternatives are often highly correlated due to overlapping paths
- Data collection
- Large size of the choice set

Dealing with correlation

Frejinger, E. and Bierlaire, M. (2007). Capturing correlation with subnetworks in route choice models, Transportation Research Part B: Methodological 41(3):363-378.

Existing Approaches

- Few models explicitly capturing correlation have been used on large-scale route choice problems
  - C-Logit (Cascetta et al., 1996)
  - Path Size Logit (Ben-Akiva and Bierlaire, 1999)
  - Link-Nested Logit (Vovsha and Bekhor, 1998)
  - Logit Kernel model adapted to route choice situation (Bekhor et al., 2002)
- Probit model (Daganzo, 1977) permits an arbitrary covariance structure specification but cannot be applied in a large-scale route choice context

Existing Approaches

- Link based path-multilevel logit model (Marzano and Papola, 2005)
  - Illustrated on simple examples and not estimated on real data

Subnetworks

How can we explicitly capture the most important correlation structure without considerably increasing the model complexity?

Subnetworks

How can we explicitly capture the most important correlation structure without considerably increasing the model complexity?

- Which are the behaviorally important decisions?

Subnetworks

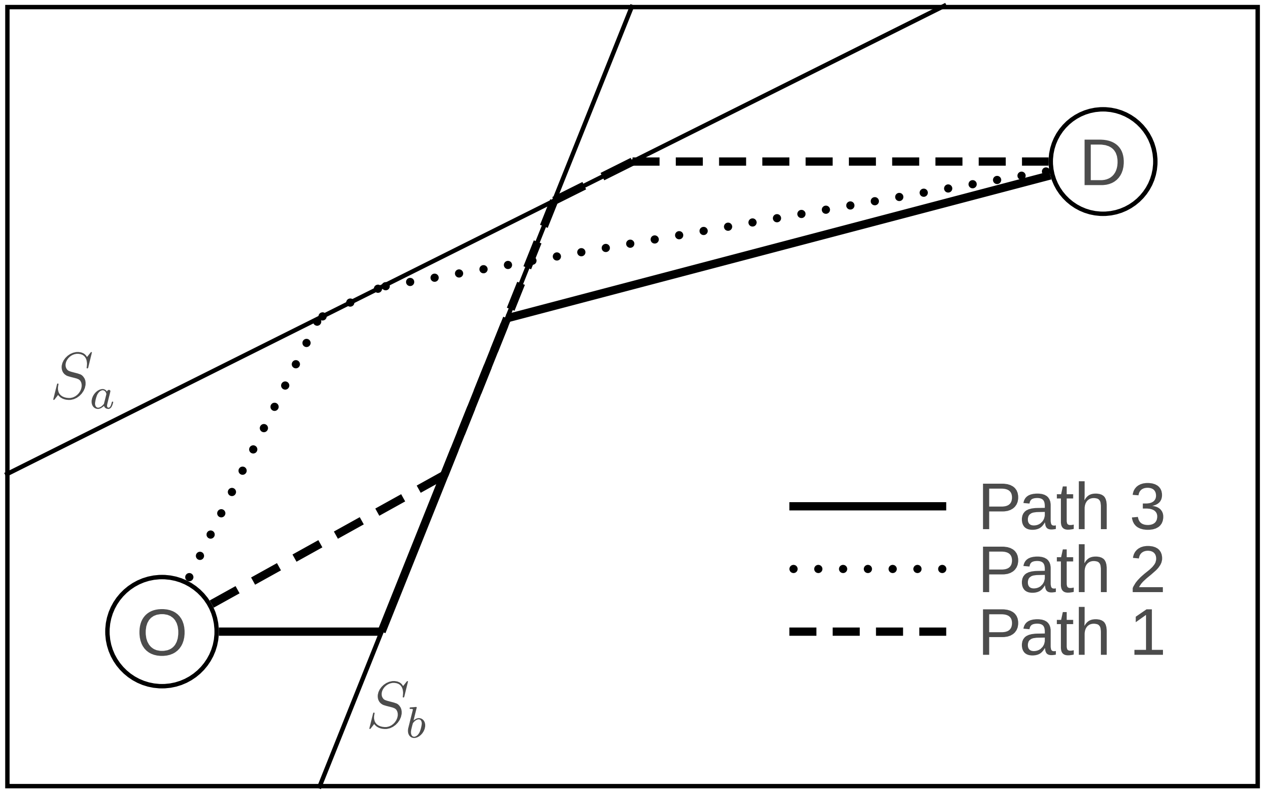
How can we explicitly capture the most important correlation structure without considerably increasing the model complexity?

- Which are the behaviorally important decisions?
- Our hypothesis: choice of specific parts of the network (e.g. main roads, city center)
- Concept: subnetwork

Subnetworks

- Subnetwork approach designed to be behaviorally realistic and convenient for the analyst
- Subnetwork component is a set of links corresponding to a part of the network which can be easily labeled
- Paths sharing a subnetwork component are assumed to be correlated even if they are not physically overlapping

Subnetworks - Example



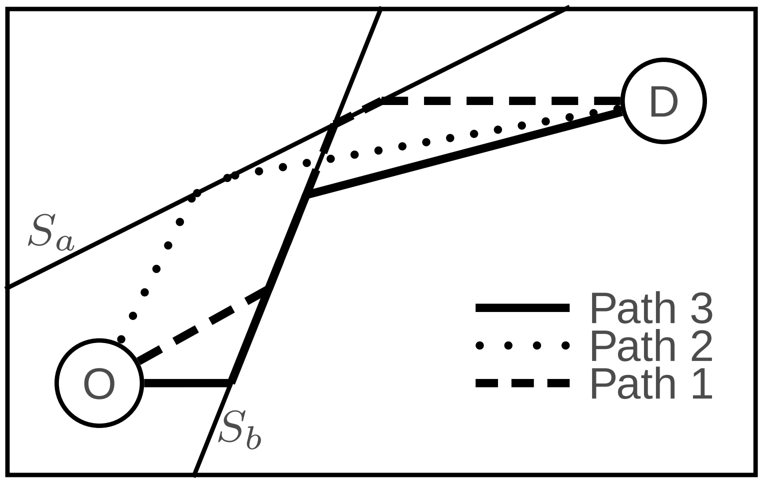
Subnetworks - Methodology

- Factor analytic specification of an error component model (based on model presented in Bekhor et al., 2002)

$$\mathbf{U}_n = \beta^T \mathbf{X}_n + \mathbf{F}_n \mathbf{T} \zeta_n + \nu_n$$

  - $\mathbf{F}_{n\ (J\text{x}Q)}$: factor loadings matrix
  - $(f_n)_{iq} = \sqrt{l_{niq}}$
  - $\mathbf{T}_{(Q\text{x}Q)} = \text{diag}\left(\sigma_1, \sigma_2, \ldots, \sigma_Q\right)$
  - $\zeta_{n\ (Q\text{x}1)}$: vector of i.i.d. N(0,1) variates
  - $\nu_{(J\text{x}1)}$: vector of i.i.d. Extreme Value distributed variates

Subnetworks - Example



$$U_1 = \beta^T X_1 + \sqrt{l_{1a}}\sigma_a\zeta_a + \sqrt{l_{1b}}\sigma_b\zeta_b + \nu_1$$

$$U_2 = \beta^T X_2 + \sqrt{l_{2a}}\sigma_a\zeta_a + \nu_2$$

$$U_3 = \beta^T X_3 + \sqrt{l_{3b}}\sigma_b\zeta_b + \nu_3$$

$$\mathbf{F}\mathbf{T}\mathbf{T}^T\mathbf{F}^T =$$

$$\begin{bmatrix} l_{1a}\sigma_a^2 + l_{1b}\sigma_b^2 & \sqrt{l_{1a}}\sqrt{l_{2a}}\sigma_a^2 & \sqrt{l_{1b}}\sqrt{l_{3b}}\sigma_b^2 \\ \sqrt{l_{1a}}\sqrt{l_{2a}}\sigma_a^2 & l_{2a}\sigma_a^2 & 0 \\ \sqrt{l_{3b}}\sqrt{l_{1b}}\sigma_b^2 & 0 & l_{3b}\sigma_b^2 \end{bmatrix}$$

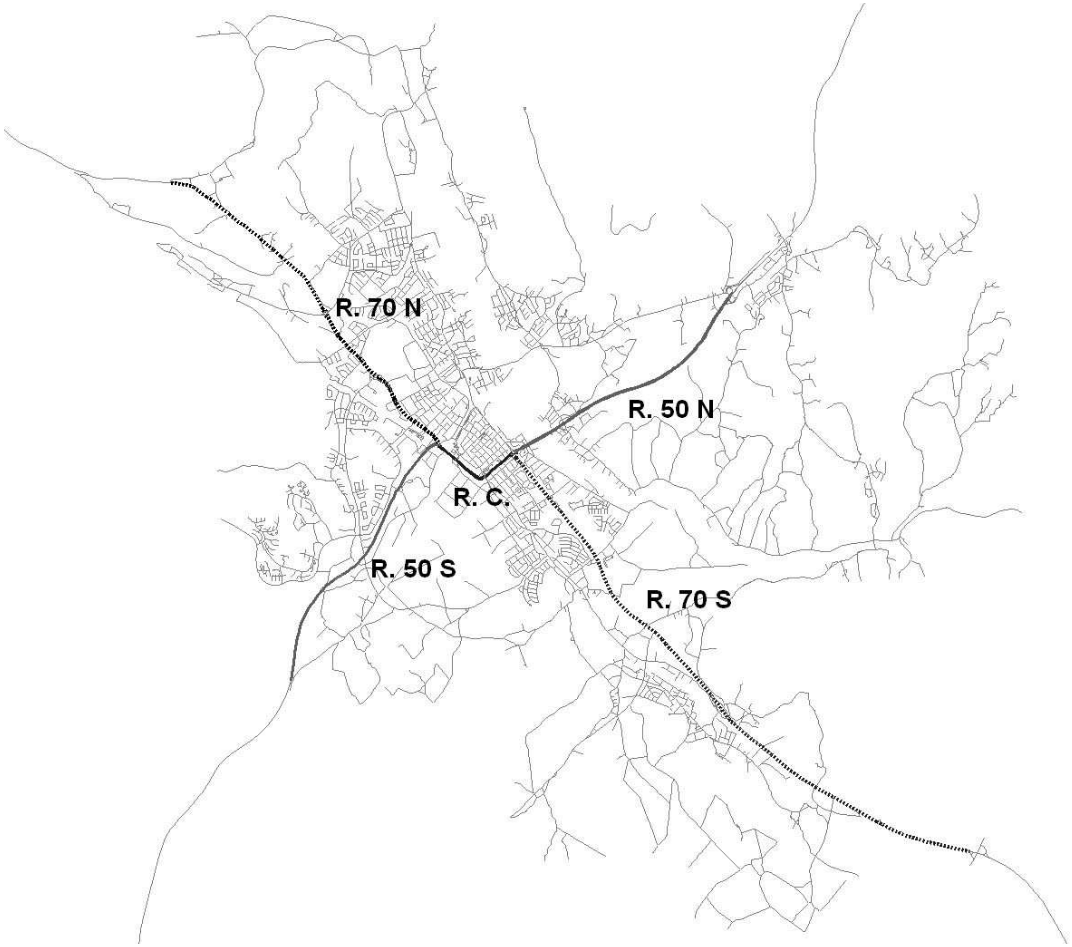
TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Empirical Results

- The approach has been tested on three datasets: Boston (Ramming, 2001), Switzerland, and Borlänge
- Deterministic choice set generation
  Link elimination
- GPS data from 24 individuals
  2978 observations, 2179 origin-destination pairs
- Borlänge network
  3077 nodes and 7459 links
- BIOGEME (biogeme.epfl.ch, Bierlaire, 2003) has been used for all model estimations

Borlänge Road Network



Model Specifications

- Six different models: MNL, PSL, EC_1, EC'_1, EC_2 and EC'_2
- EC_1 and EC'_1 have a simplified correlation structure
- EC'_1 and EC'_2 do not include a Path Size attribute
- Deterministic part of the utility

$$
\begin{aligned}
V_i = \ & \beta_{\text{PS}} \ln(\text{PS}_i) + \beta_{\text{EstimatedTime}} \text{EstimatedTime}_i + \\
& \beta_{\text{NbSpeedBumps}} \text{NbSpeedBumps}_i + \beta_{\text{NbLeftTurns}} \text{NbLeftTurns}_i + \\
& \beta_{\text{AvgLinkLength}} \text{AvgLinkLength}_i
\end{aligned}
$$

Estimation Results

- Parameter estimates for explanatory variables are stable across the different models
- Path size parameter estimates

Parameter	PSL	EC_1	EC_2
Path Size	-0.28	-0.49	-0.53
Scaled estimate	-0.33	-0.53	-0.56
Rob. T-test 0	-4.05	-5.61	-5.91

- All covariance parameters estimates in the different models are significant except the one associated with R.50 S

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Estimation Results

Model	Nb. σ Estimates	Nb. Estimated Parameters	Final L-L	Adjusted Rho-Square
MNL	-	12	-4186.07	0.152
PSL	-	13	-4174.72	0.154
EC_1 (with PS)	1	14	-4142.40	0.161
EC'_1	1	13	-4165.59	0.156
EC_2 (with PS)	5	18	-4136.92	0.161
EC'_2	5	17	-4162.74	0.156

1000 pseudo-random draws for Maximum Simulated Likelihood estimation

2978 observations

Null log likelihood: -4951.11

BIOGEME (biogeme.epfl.ch) has been used for all model estimations.

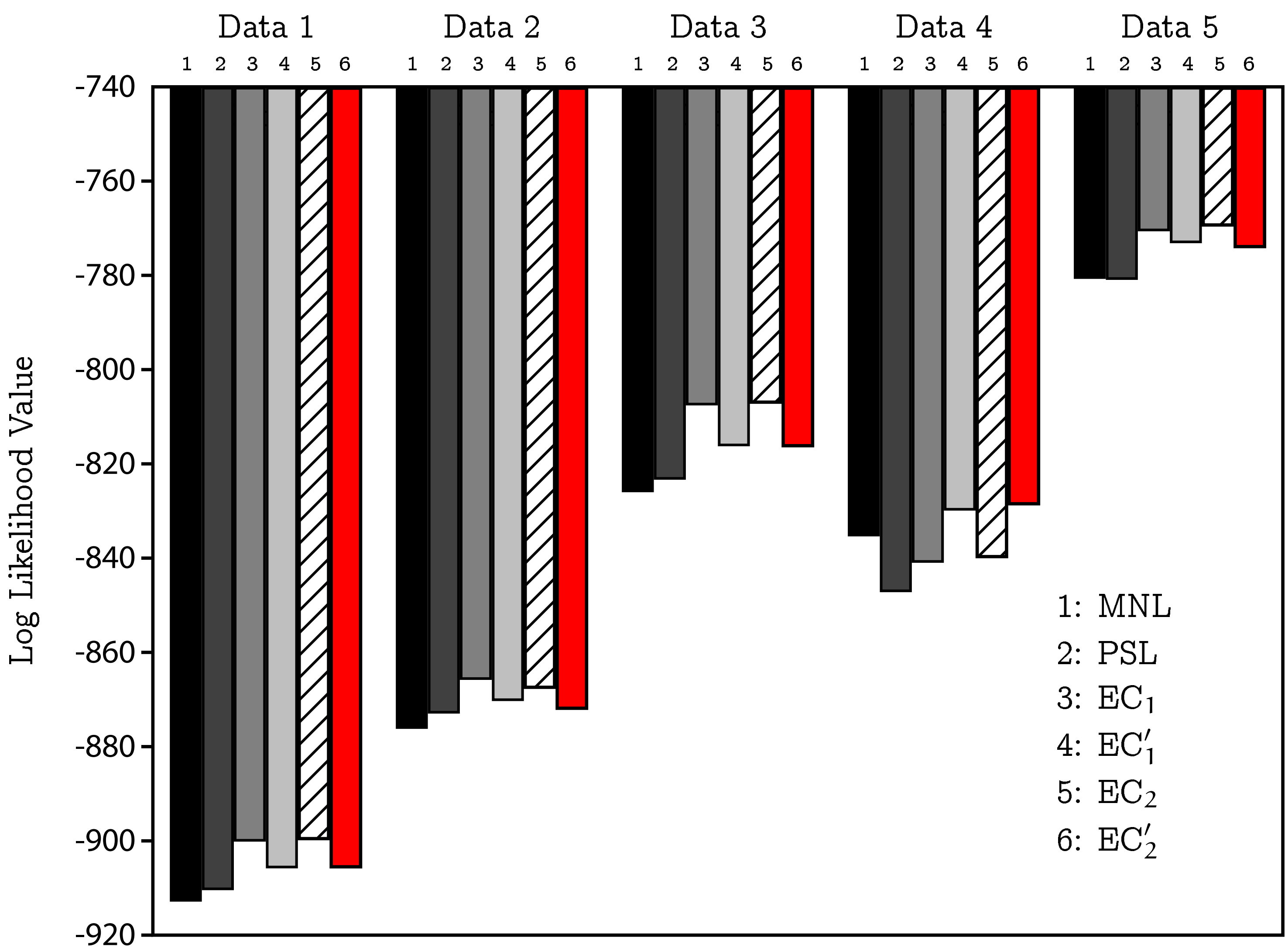
Forecasting Results

- Comparison of the different models in terms of their performance of predicting choice probabilities
- Five subsamples of the dataset
  - Observations corresponding to 80% of the origin destination pairs (randomly chosen) are used for estimating the models
  - The models are applied on the observations corresponding to the other 20% of the origin destination pairs
- Comparison of final log-likelihood values

Forecasting Results

- Same specification of deterministic utility function for all models
- Same interpretation of these models as for those estimated on the complete dataset
- Coefficient and covariance parameter values are stable across models

Forecasting Results



Conclusion - Subnetworks

- Models based on subnetworks are designed for route choice modeling of realistic size
- Correlation on subnetwork is explicitly captured within a factor analytic specification of an Error Component model
- Estimation and prediction results clearly shows the superiority of the Error Component models compared to PSL and MNL
- The subnetwork approach is flexible and the model complexity can be controlled by the analyst

Network-free data

- Bierlaire, M., Frejinger, E., and Stojanovic, J. (2006). A latent route choice model in Switzerland. Proceedings of the European Transport Conference (ETC) September 18-20, 2006.
- Bierlaire, M., and Frejinger, E. (2007). Route choice modeling with network-free data. Technical report TRANSP-OR 070214. Transport and Mobility Laboratory, ENAC, EPFL.

Data collection and processing

- Link-by-link descriptions of chosen routes necessary for route choice modeling but never directly available
- Data processing in order to obtain network compliant paths
  - Map matching of GPS points
  - Reconstruction of reported paths
- Difficult to verify and may introduce bias and errors

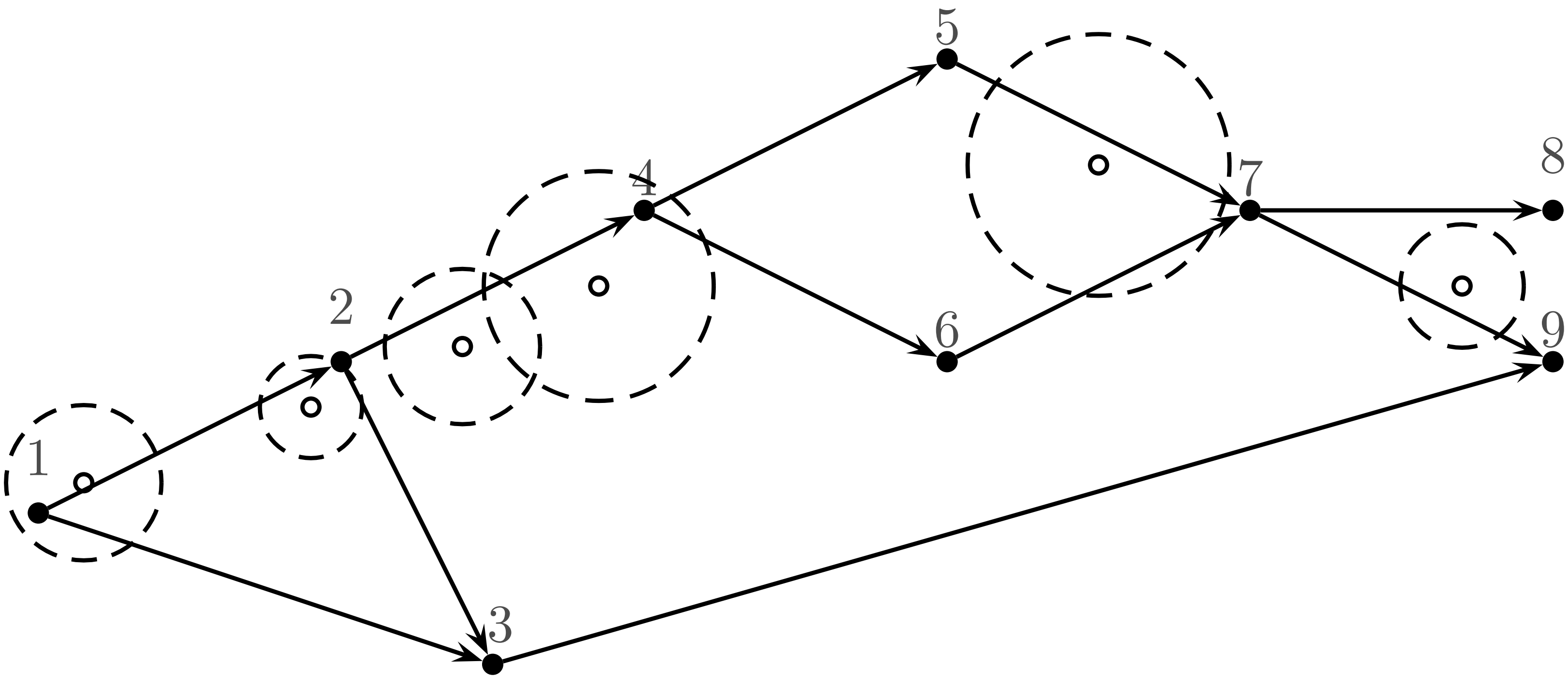
Modeling with network-free data

- An observation i is a sequence of individual pieces of data related to an itinerary. Examples: sequence of GPS points or reported locations
- For each piece of data we define a Domain of Data Relevance (DDR) that is the physical area where it is relevant
- The DDRs bridge the gap between the network-free data and the network model

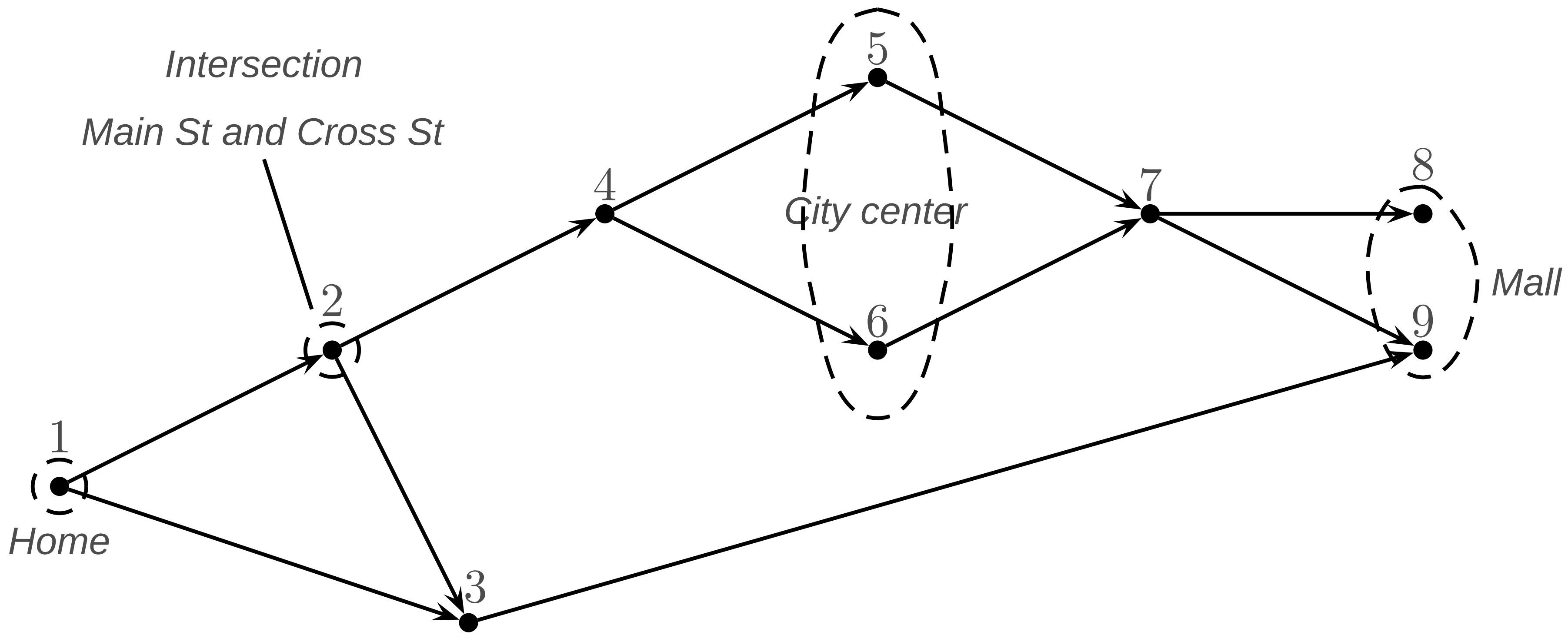
TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Example - GPS data



Example - Reported trip



Domain of Data Relevance

- For each piece of data d we generate a list of relevant network elements e (links and nodes)
  We define an indicator function

$$\delta(d,e) = \begin{cases} 1 & \text{if } e \text{ is related to the DDR of } d \\ 0 & \text{otherwise} \end{cases}$$

Model estimation

- We aim at estimating the parameters β of route choice model $P(p|\mathcal{C}_n(s);\beta)$
- We have a set $\mathcal{S}_i$ of relevant od pairs
- The probability of reproducing observation i of traveler n, given $\mathcal{S}_i$ is defined as

$$P_n(i|\mathcal{S}_i) = \sum_{s\in\mathcal{S}_i} P_n(s|\mathcal{S}_i) \sum_{p\in\mathcal{C}_n(s)} P_n(i|p) P_n(p|\mathcal{C}_n(s);\beta)$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Model estimation

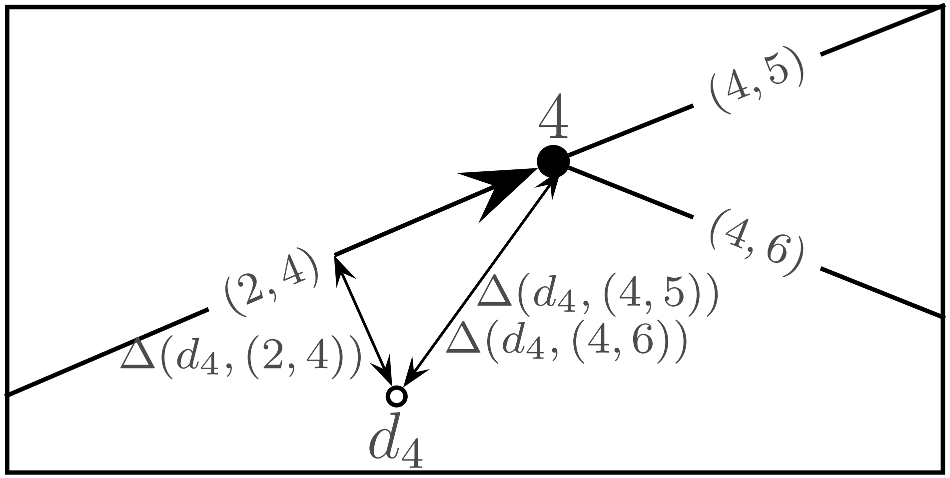
- Measurement equation $P_n(i|p)$
  - Reported trips

$$P_n(i|p) = \begin{cases} 1 & \text{if } i \text{ corresponds to } p \\ 0 & \text{otherwise} \end{cases}$$

  - GPS data
    $P_n(i|p) = 0$ if i does not correspond to p
    If i corresponds to p then $P_n(i|p)$ is a function of the distance between i and p

Model estimation

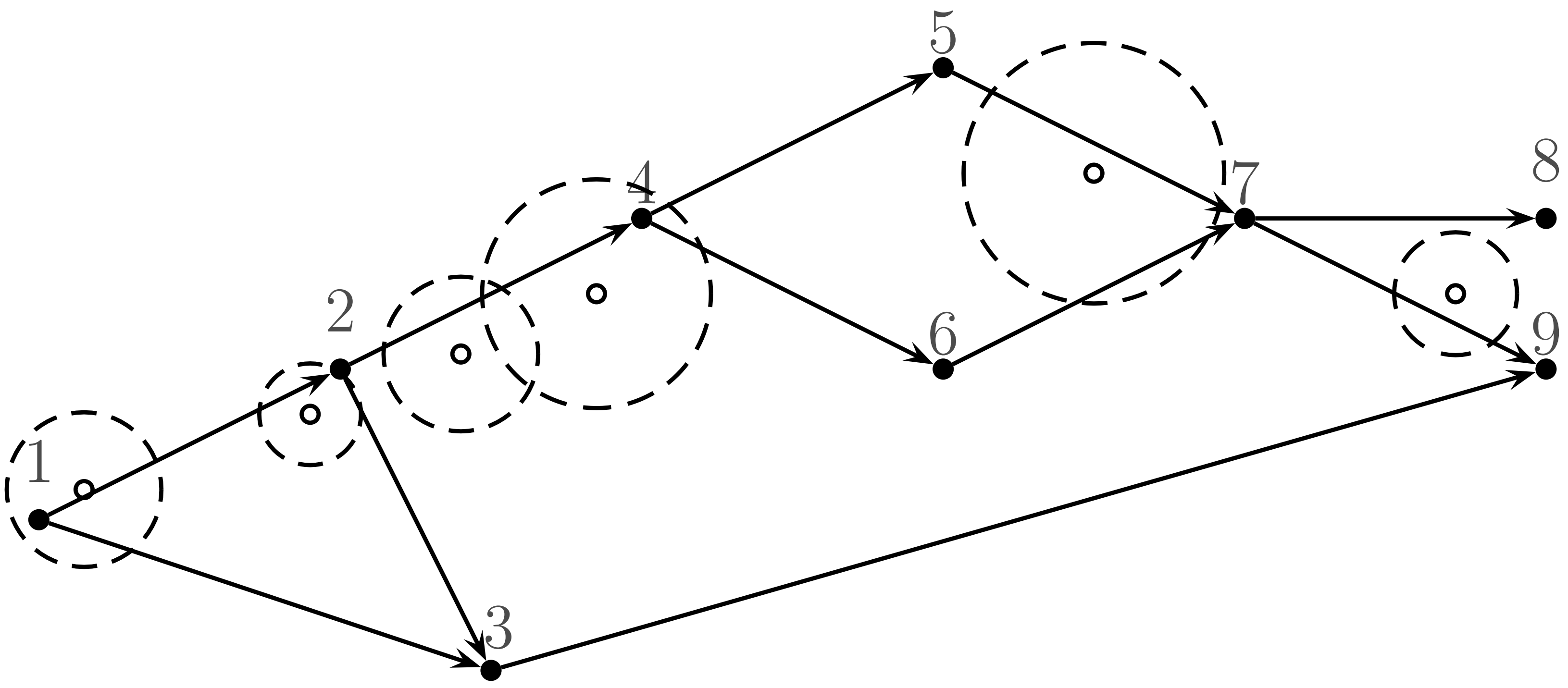
- Measurement equation $P_n(i|p)$ for GPS data
- Distance between i and a the closest point on a link ℓ is $D(d,p) = \min_{\ell \in A_{pd}} \Delta(d,\ell)$



TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Model estimation



$$P_n(i|\mathcal{S}_i) = \sum_{s \in \mathcal{S}_i} P_n(s|\mathcal{S}_i) \sum_{p \in \mathcal{C}_n(s)} P_n(i|p) P_n(p|\mathcal{C}_n(s); \beta)$$

$$P(i|s) = P(i|p_1)P(p_1|\mathcal{C}(s); \beta) + P(i|p_2)P(p_2|\mathcal{C}(s); \beta)$$

Empirical Results

- Simplified Swiss network (39411 links and 14841 nodes)
- RP data collection through telephone interviews
- Long distance car travel
- The chosen routes are described with the origin and destination cities as well as 1 to 3 cities or locations that the route pass by
- 940 observations available after data cleaning and verification

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Empirical Results

Empirical Results

- No information available on the exact origin destination pairs

$$P(s|i) = \frac{1}{|S_i|} \; \forall s \in S_i$$

- $P(r|i)$ is modeled with a binary variable

$$\delta_{ri} = \begin{cases} 1 & \text{if } r \text{ corresponds to } i \\ 0 & \text{otherwise} \end{cases}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Empirical Results

- Two origin-destination pairs are randomly chosen for each observation
- 46 routes per choice set are generated with a choice set generation algorithm
- After choice set generation 780 observations are available
  - 160 observations were removed because either all or none of the generated routes crossed the observed zones

Empirical Results

- Probability of an aggregate observation i

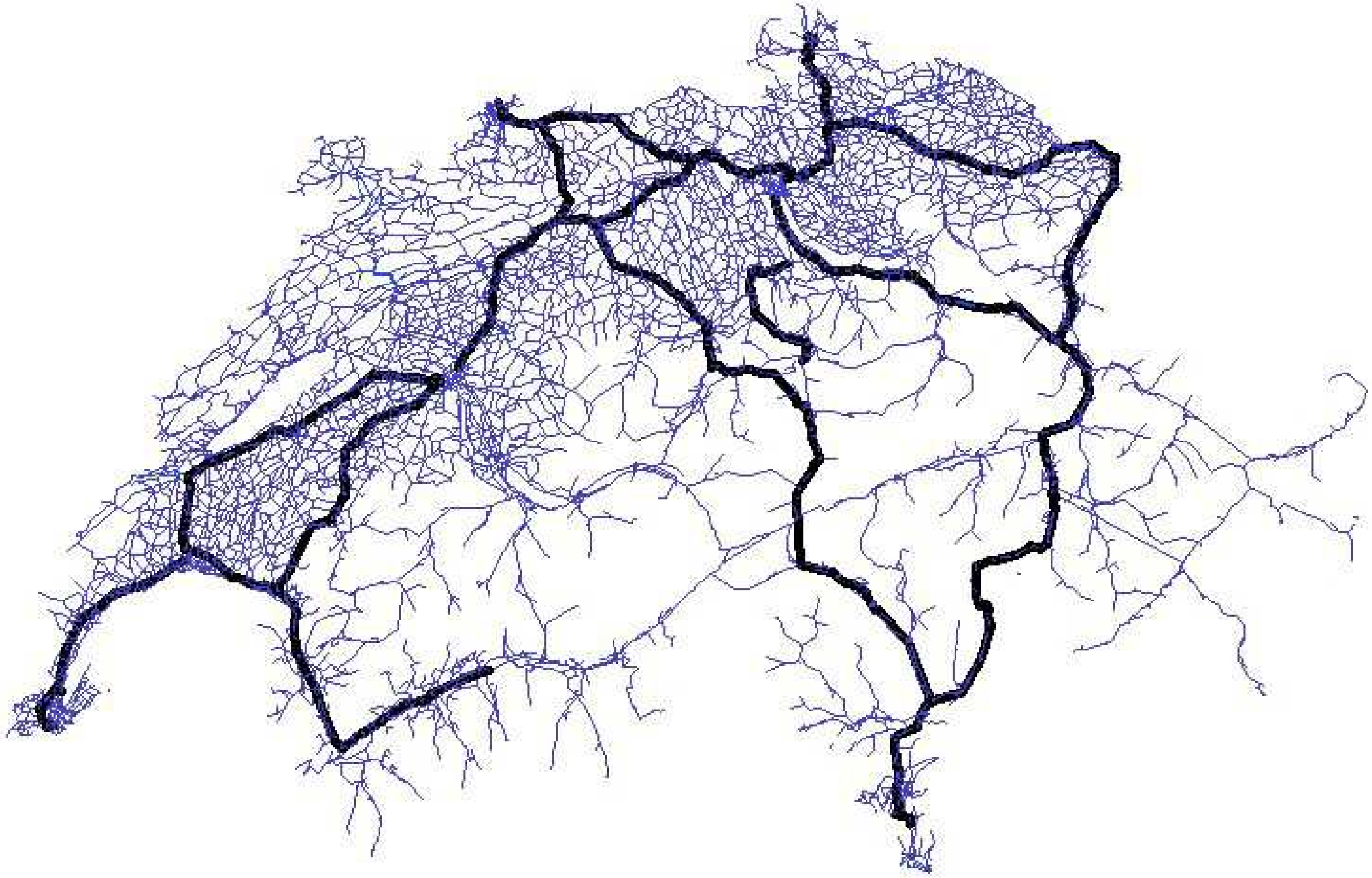
$$P(i) = \sum_{s \in S_i} \frac{1}{|S_i|} \sum_{r \in C_s} \delta_{ri} P(r|C_s)$$

- We estimate Path Size Logit (Ben-Akiva and Bierlaire, 1999) and Subnetwork (Frejinger and Bierlaire, 2007) models
- BIOGEME (biogeme.epfl.ch) used for all model estimations

Empirical Results - Subnetwork

- Subnetwork: main motorways in Switzerland
- Correlation among routes is explicitly modeled on the subnetwork
- Combined with a Path Size attribute
- Linear-in-parameters utility specifications

Empirical Results - Subnetwork



Parameter	PSL		Subnetwork	
ln(path size) based on free-flow time	**1.04**	(0.134) 7.81	**1.10**	(0.141) 7.78
Scaled Estimate	*1.04*		*1.04*	
Freeway free-flow time 0-30 min	**-7.12**	(0.877) -8.12	**-7.45**	(0.984) -7.57
Scaled Estimate	*-7.12*		*-7.04*	
Freeway free-flow time 30min - 1 hour	**-1.69**	(0.875) -1.93	**-2.26**	(1.03) -2.19
Scaled Estimate	*-1.69*		*-2.14*	
Freeway free-flow time 1 hour +	**-4.98**	(0.772) -6.45	**-5.64**	(1.00) -5.61
Scaled Estimate	*-4.98*		*-5.33*	
CN free-flow time 0-30 min	**-6.03**	(0.882) -6.84	**-6.25**	(0.975) -6.41
Scaled Estimate	*-6.03*		*-5.91*	
CN free-flow time 30 min +	**-1.87**	(0.331) -5.64	**-2.16**	(0.384) -5.63
Scaled Estimate	*-1.87*		*-2.04*	
Main free-flow travel time 10 min +	**-2.03**	(0.502) -4.05	**-2.46**	(0.624) -3.95
Scaled Estimate	*-2.03*		*-2.33*	
Small free-flow travel time	**-2.16**	(0.685) -3.16	**-2.75**	(0.804) -3.42
Scaled Estimate	*-2.16*		*-2.60*	
Proportion of time on freeways	**-2.2**	(0.812) -2.71	**-2.31**	(0.865) -2.67
Scaled Estimate	*-2.2*		*-2.18*	
Proportion of time on CN	**0 fixed**		**0 fixed**	
Proportion of time on main	**-4.43**	(0.752) -5.88	**-4.40**	(0.800) -5.51
Scaled Estimate	*-4.43*		*-4.16*	
Proportion of time on small	**-6.23**	(0.992) -6.28	**-6.02**	(1.03) -5.83
Scaled Estimate	*-6.23*		*-5.69*	
Covariance parameter			**0.217**	(0.0543) 4.00
Scaled Estimate			*0.205*	

Empirical Results

	PSL	Subnetwork
Covariance parameter		0.217
(Rob. Std. Error) Rob. T-test		(0.0543) 4.00
Number of simulation draws	-	1000
Number of parameters	11	12
Final log-likelihood	-1164.850	-1161.472
Adjusted rho square	0.145	0.147
Sample size: 780, Null log-likelihood: -1375.851		

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Empirical Results

- All parameters have their expected signs and are significantly different from zero
- The values and significance level are stable across the two models
- The subnetwork model is significantly better than the Path Size Logit (PSL) model

Concluding remarks

- Network-free data are more reliable
- Data processing may bias the result
- We prefer to model explicitly the relationship between the data and the model

Choice set generation

Frejinger, E. and Bierlaire, M. (2007). Stochastic Path Generation Algorithm for Route Choice Models. Proceedings of the Sixth Triennial Symposium on Transportation Analysis (TRISTAN) June 10-15, 2007.

Introduction

- Choice sets need to be defined prior to the route choice modeling
- Path enumeration algorithms are used for this purpose, many heuristics have been proposed, for example:
  - Deterministic approaches: link elimination (Azevedo et al., 1993), labeled paths (Ben-Akiva et al., 1984)
  - Stochastic approaches: simulation (Ramming, 2001) and doubly stochastic (Bovy and Fiorenzo-Catalano, 2006)

Introduction

- Underlying assumption: the actual choice set is generated
- Empirical results suggest that this is not always true
- Our approach:
    - True choice set = universal set
    - Too large
    - Sampling of alternatives

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Sampling of Alternatives

- Multinomial logit model (e.g. Ben-Akiva and Lerman, 1985):

$$P(i|\mathcal{C}_n) = \frac{q(\mathcal{C}_n|i)P(i)}{\sum_{j\in\mathcal{C}_n} q(\mathcal{C}_n|j)P(j)} = \frac{e^{V_{in}+\ln q(\mathcal{C}_n|i)}}{\sum_{j\in\mathcal{C}_n} e^{V_{jn}+\ln q(\mathcal{C}_n|j)}}$$

$\mathcal{C}_n$: set of sampled alternatives
$q(\mathcal{C}_n|j)$: probability of sampling $\mathcal{C}_n$ given that j is the chosen alternative

Importance Sampling of Alternatives

- Attractive paths have higher probability of being sampled than unattractive paths
- Path utilities must be corrected in order to obtain unbiased estimation results

Stochastic Path Enumeration

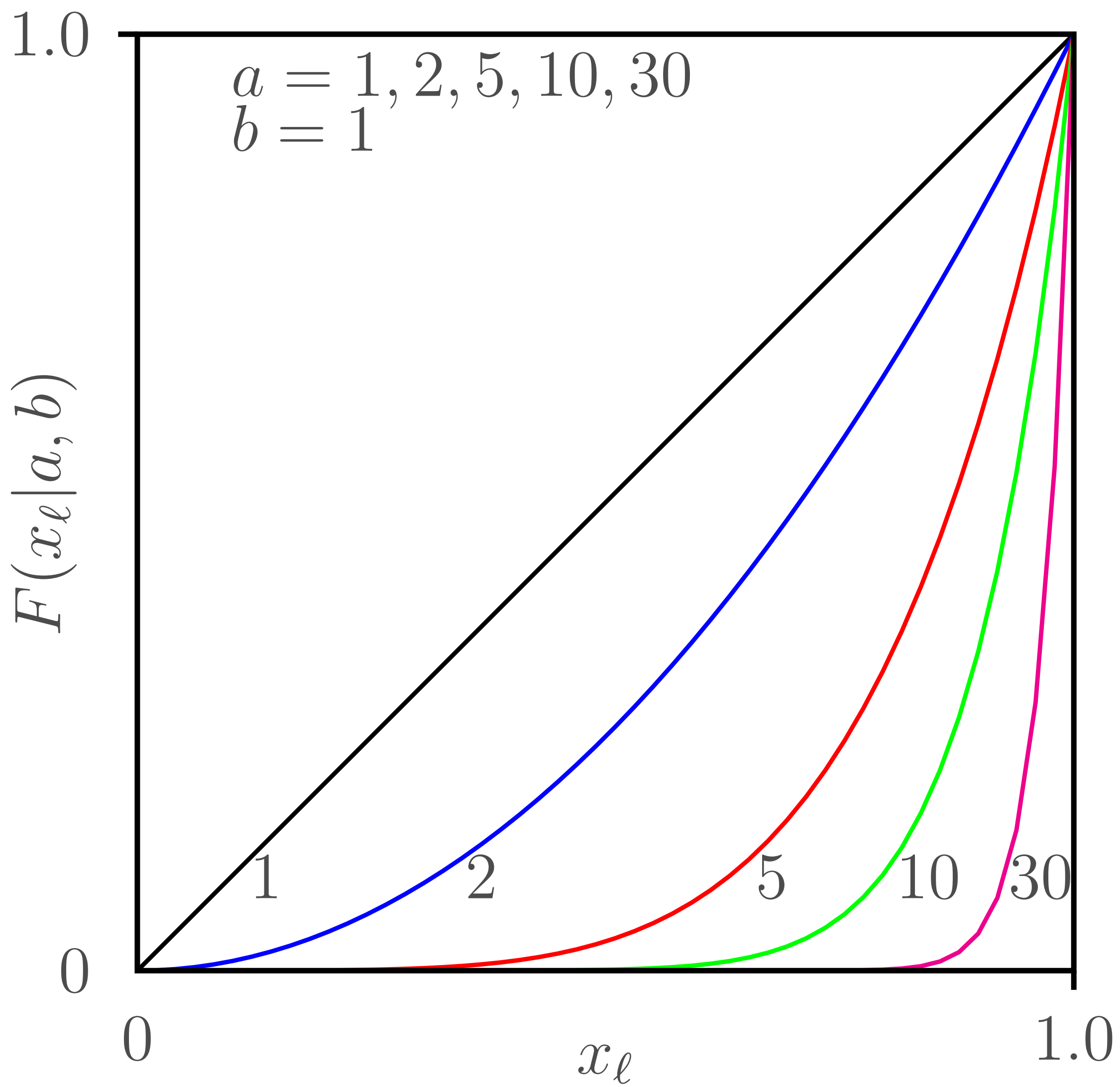
- Flexible approach that can be combined with various algorithms, here a biased random walk approach
- The probability of a link ℓ with source node v and sink node w is modeled in a stochastic way based on its distance to the shortest path
- Kumaraswamy distribution, cumulative distribution function $F(x_\ell | a, b) = 1 - (1 - {x_\ell}^a)^b$ for $x_\ell \in [0, 1]$.

$$x_\ell = \frac{SP(v,d)}{C(\ell) + SP(w,d)}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Stochastic Path Enumeration



Stochastic Path Enumeration

- Probability for path j to be sampled

$$q(j) = \prod_{\ell=(v,w)\in\Gamma_j} q((v,w)|\mathcal{E}_v)$$

- Γ_j: ordered set of all links in j
- v: source node of j
- $\mathcal{E}_v$: set of all outgoing links from v
- Issue: in theory, the set of all paths $\mathcal{U}$ is unbounded. We treat it as bounded with size J.

Sampling of Alternatives

- Following Ben-Akiva (1993)
- Sampling protocol
  1. A set $\widetilde{\mathcal{C}}_n$ is generated by drawing R paths with replacement from the universal set of paths $\mathcal{U}$
  2. Add chosen path to $\widetilde{\mathcal{C}}_n$
- Outcome of sampling: $(\widetilde{k}_1, \widetilde{k}_2, \ldots, \widetilde{k}_J)$ and $\sum_{j=1}^{J} \widetilde{k}_j = R$

$$P(\widetilde{k}_1, \widetilde{k}_2, \ldots, \widetilde{k}_J) = \frac{R!}{\prod_{j \in \mathcal{U}} \widetilde{k}_j!} \prod_{j \in \mathcal{U}} q(j)^{\widetilde{k}_j}$$

- Alternative j appears $k_j = \widetilde{k}_j + \delta_{cj}$ in $\widetilde{\mathcal{C}}_n$

Sampling of Alternatives

- Let $\mathcal{C}_n = \{j \in \mathcal{U} \mid k_j > 0\}$

$$q(\mathcal{C}_n|i) = q(\widetilde{\mathcal{C}}_n|i) = \frac{R!}{(k_i - 1)! \prod_{\substack{j\in\mathcal{C}_n \\ j\neq i}} k_j!} q(i)^{k_i-1} \prod_{\substack{j\in\mathcal{C}_n \\ j\neq i}} q(j)^{k_j} = K_{\mathcal{C}_n} \frac{k_i}{q(i)}$$

$$K_{\mathcal{C}_n} = \frac{R!}{\prod_{j\in\mathcal{C}_n} k_j!} \prod_{j\in\mathcal{C}_n} q(j)^{k_j}$$

$$P(i|\mathcal{C}_n) = \frac{e^{V_{in} + \ln\left(\frac{k_i}{q(i)}\right)}}{\sum_{j\in\mathcal{C}_n} e^{V_{jn} + \ln\left(\frac{k_j}{q(j)}\right)}}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

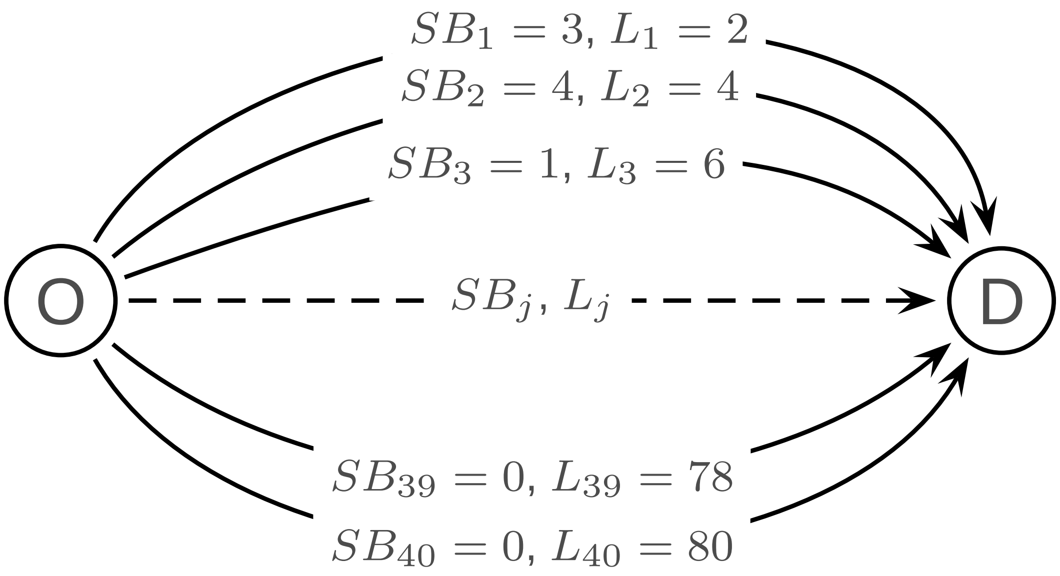
Preliminary Numerical Results

- Estimation of models based on synthetic data generated with postulated models
  - Non-correlated paths
    Postulated model same as estimated model (multinomial logit)
  - Correlated paths in a “grid-like” network
    Postulated model is probit and estimated models are multinomial logit and path size logit
- True parameter values are compared to estimates

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Preliminary Numerical Results



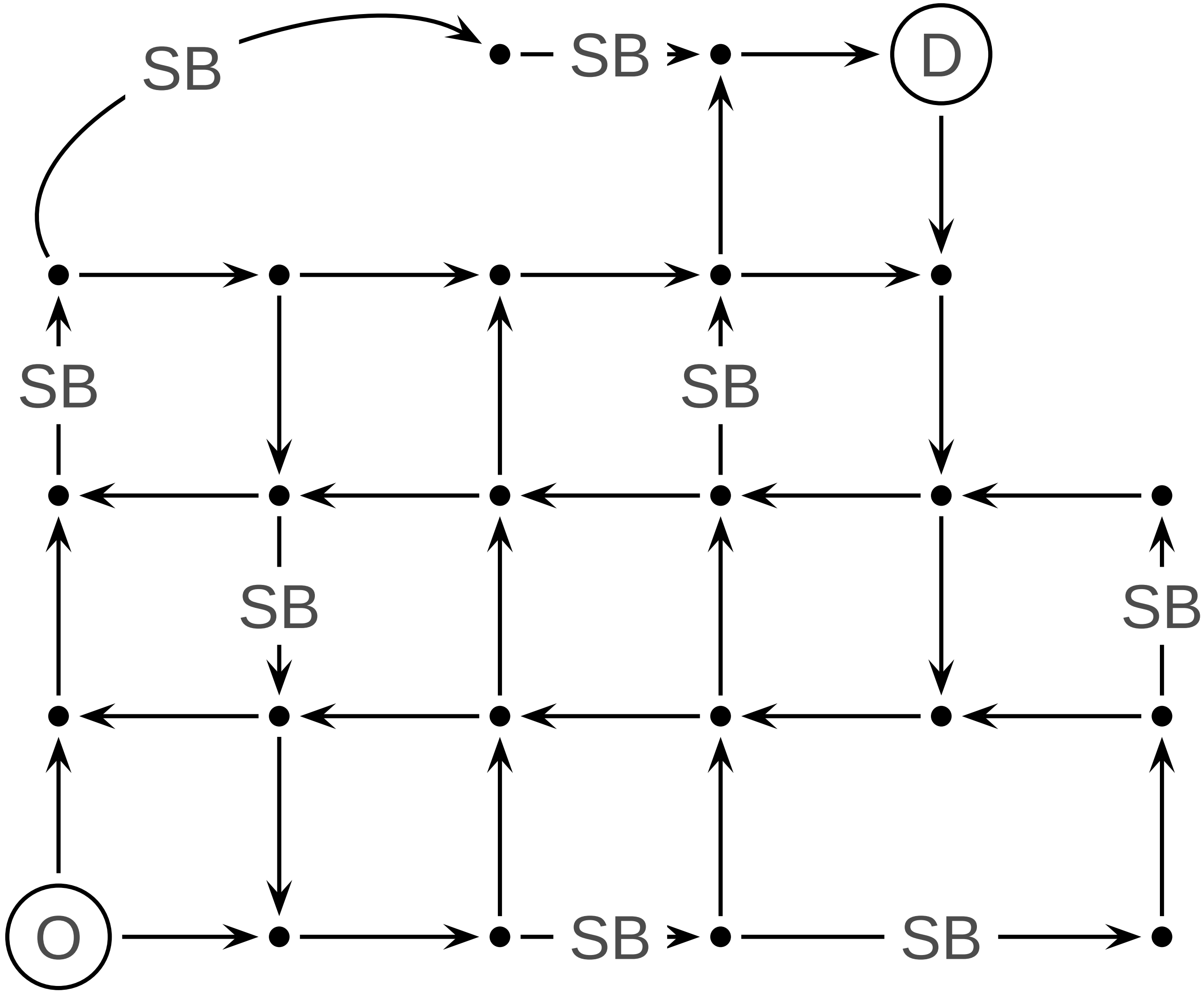
Preliminary Numerical Results

- True model: multinomial logit
  $U_j = \beta_{\text{L}}\ \text{length}_j + \beta_{\text{SB}}\ \text{nbspeedbumps}_j + \varepsilon_j$
  $\beta_{\text{L}} = -0.6$ and $\beta_{\text{SB}} = -0.3$
  ε_j is distributed Extreme Value with location parameter 0 and scale 1
- 500 observations, therefore 500 choice sets are sampled
- Biased random walk using 40 draws with $a = 2$ and $b = 1$
  Generated choice sets include at least 7, maximum 18 and on average 11.9 paths

Preliminary Numerical Results

Sampling correction	MNL without	MNL with
$\widehat{\beta}_{\mathrm{L}}$ (-0.6)	-0.203	-0.286
Scaled estimate	-0.600	-0.600
Robust std.	0.0193	0.019
Robust t-test	-10.53	-15.01
$\widehat{\beta}_{\mathrm{SB}}$ (-0.3)	-0.0194	-0.143
Scaled estimate	-0.0573	-0.300
Robust std.	0.0662	0.0661
Robust t-test	-0.29	-2.17
Null log-likelihood	-1069.453	-1633.501
Final log-likelihood	-788.42	-759.848
Adjusted $\bar{\rho}^2$	0.261	0.288
BIOGEME has been used for all model estimations.		

Preliminary Numerical Results



Preliminary Numerical Results

- True model: probit (Burrell, 1968)
  $U_\ell = \beta_{\text{L}}$ length$_\ell + \beta_{\text{SB}}$ nbspeedbumps$_\ell + \sigma\sqrt{L_\ell}\nu_\ell$
  $\beta_{\text{L}} = -0.6$ and $\beta_{\text{SB}} = -0.4$
  ν_ℓ is distributed standard Normal
  Link utility variance assumed proportional to length with parameter $\sigma = 0.8$
- Path utilities are link additive
- 382 observations are generated after 500 realizations of the link utilities

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Preliminary Numerical Results

- Biased random walk using 30 draws with $a = 2$ and $b = 1$ (382 choice sets)
  Generated choice sets include at least 7, maximum 19 and on average 13.5 paths

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Preliminary Numerical Results

Sampling correction	MNL without	MNL with	PSL without	PSL with
$\widehat{\beta}_L$ (-0.6)	-0.627	-0.978	-0.619	-0.969
Scaled estimate	-0.600	-0.600	-0.600	-0.600
Robust std.	0.0397	0.032	0.0407	0.0358
Robust t-test	-15.79	-30.57	-15.22	-27.04
$\widehat{\beta}_{SB}$ (-0.4)	-0.0822	-0.0801	-0.347	-0.461
Scaled estimate	-0.0787	-0.0491	-0.336	-0.285
Robust std.	0.052	0.0559	0.182	0.158
Robust t-test	-1.58	-1.43	-1.90	-2.92
$\widehat{\beta}_{PS}$			1.17	1.74
Scaled estimate			1.13	1.08
Robust std.			0.788	0.705
Robust t-test			1.49	2.47

Preliminary Numerical Results

Sampling correction	MNL without	MNL with	PSL without	PSL with
Null log-likelihood	-988.63	-2769.959	-988.63	-2769.959
Final log-likelihood	-676.111	-653.396	-674.481	-649.268
Adjusted $\bar{\rho}^2$	0.314	0.337	0.315	0.340
BIOGEME has been used for all model estimations.				

Conclusions and Future Work

- Stochastic path enumeration algorithms are viewed as an approach for importance sampling of alternatives
- We propose an algorithm that allows for computation of path selection probabilities and correction for sampling
- Ongoing research, further work will be dedicated, for example, to empirical results on real data and correction in prediction