Preferred citation style for this presentation

K.W. Axhausen (2005) Data collection methods and models for consumer choice behaviour: Examples from the transport sector, CEPE Kolloqium, Zürich, June 2005.

Data collection methods and models for consumer choice behaviour: Examples from the transport sector

KW Axhausen

IVT
ETH
Zürich

June 2005

Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Motivation

Calculation of equilibrium points after some policy change using

$$k = f(d(k, M, P), I, S)$$

with

Demand model $d = g(k, M, P)$ and

Supply model $k = f(d, I, S)$

k	Generalised costs	d	Demand
M	Market structure	P	„Population“
I	Infrastructure	S	Services

Motivation

Keeping market structure and population constant, we are searching for

$$d \quad = f(k)$$

or

$$d_j \quad = f(k_j)$$

$$= f(\Sigma\, ß_{j,q,t} * \; x_{j,q,t} + ß_{q,t} * c_{q,t} * x_{j,q,t} + ß_q * p_q * x_{j,q,t})$$

Focus

Discrete, disaggregate choices (for agent-based simulations)

Issues:

- data
- (one possible) statistical model form
- and an example

Types of data

Four classes:

- Properties of the person q and context (p_q)
- Revealed behaviour (d_j), description of the alternative ($x_{j,q,t}$) and its choice context ($c_{q,t}$)
- Stated responses in hypothetical markets
- Attitudes, beliefs and values

Stated response surveys (adapted from Lee-Gosselin)

	Values of context variables	
	Given	Not (all) given
Alternatives		
Given	**Stated preference** (Stated choice) (Trade-off between alternatives)	Stated tolerance (Transfer price) (Switch in alternatives)
Not given	Stated adaptation (Constructing alternatives)	Stated prospect (Learning)

Observing choices, contexts and markets

Based on our hypothesis about the variables influencing we obtain:

In revealed preference (RP) surveys:

- 1 to 3 or 4 choices without (much) control over the context variables

In stated preference (SP) surveys:

- 12-15 (and more) choices with full control over the context variables

RP: Plus and minus

Plus:

- It has happened
- Reflects current market situation

Minus:

- Not all information is available for the analyst (incomplete hypotheses before the data collection)
- Often strong correlations between variables of interest
- Too small variance of the variables of interest
- The effects of some variables of interest are too weak to be identified in the typical survey
- Difficulties in imputing the variables describing the non-chosen alternatives

SP: Plus and Minus

Plus:

- Full control over size, structure and values of the variance-covariance matrix
- Choice of variables (weak and strong)
- All information which the respondent uses is known

Minus:

- It has not happened in real life
- Inertia effects
- Limits on the number of variables
- Framing effects

Current best practise in SP construction

- Clear hypothesis
- Clear a-priori market segmentation
- Clear definition of the choice context

- Recruitment from RP surveys
- Construction of SP around an observed choice
- Efficient factorial designs
- Limit the number of experiments to 12-15 relevant ones (of two, three types) involving up to 10 variables across alternatives
- Attention to the distribution of the trade-offs offered
- Hierarchical sets of experiments for very complex choice situations

Example: *stated preference*

Ihr Busangebot	
Reisezeit:	30 min
Fahrkomfort:	gut
Takt:	alle 15 min
Zuverlässigkeit:	gut

Wie beurteilen Sie dieses Angebot ?

sehr gut						sehr schlecht
1	2	3	4	5	6	7

Example: *stated choice*

Möglichkeit 1	**Möglichkeit 2**
Sie fahren mit dem Auto	Sie fahren mit dem Zug
	Zugangszeit (von zu Hause/Ausgangsort zum Bahnhof): **15 Minuten**
Fahrzeit (Tür zu Tür) : **35 Minuten**	Fahrzeit (Zeit im System) : **20 Minuten**
	Umsteigen: **2 mal**
	Intervall (Fahrplantakt): **15 Minuten**
	Komfort: **ICN**
Preis (Reisekosten): **5 Fr.**	Preis (Reisekosten): **6 Fr.**
Wahrscheinlichkeit für eine mindestens 10-Min. Verspätung ist: **20%**	Wahrscheinlichkeit für eine mindestens 10-Min. Verspätung ist: **5%**

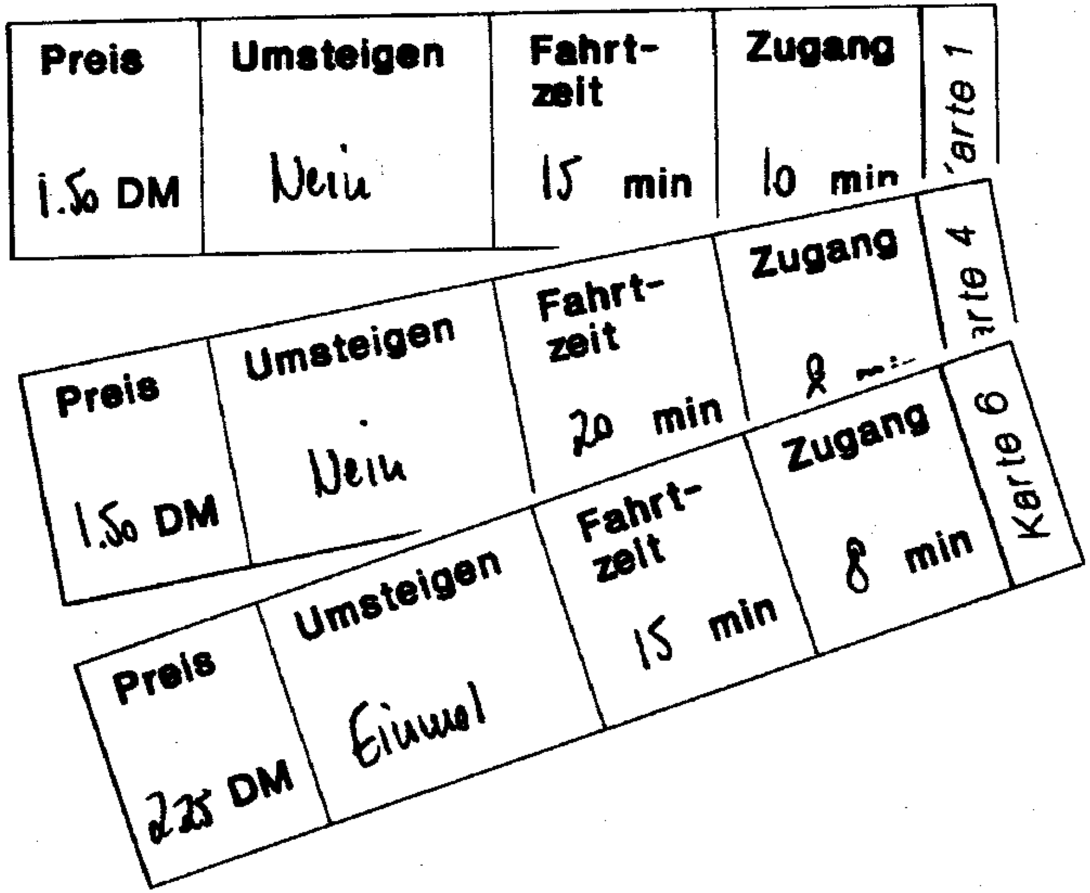
Ihre Wahl ? ❑ ❑

Example: *stated-choice*

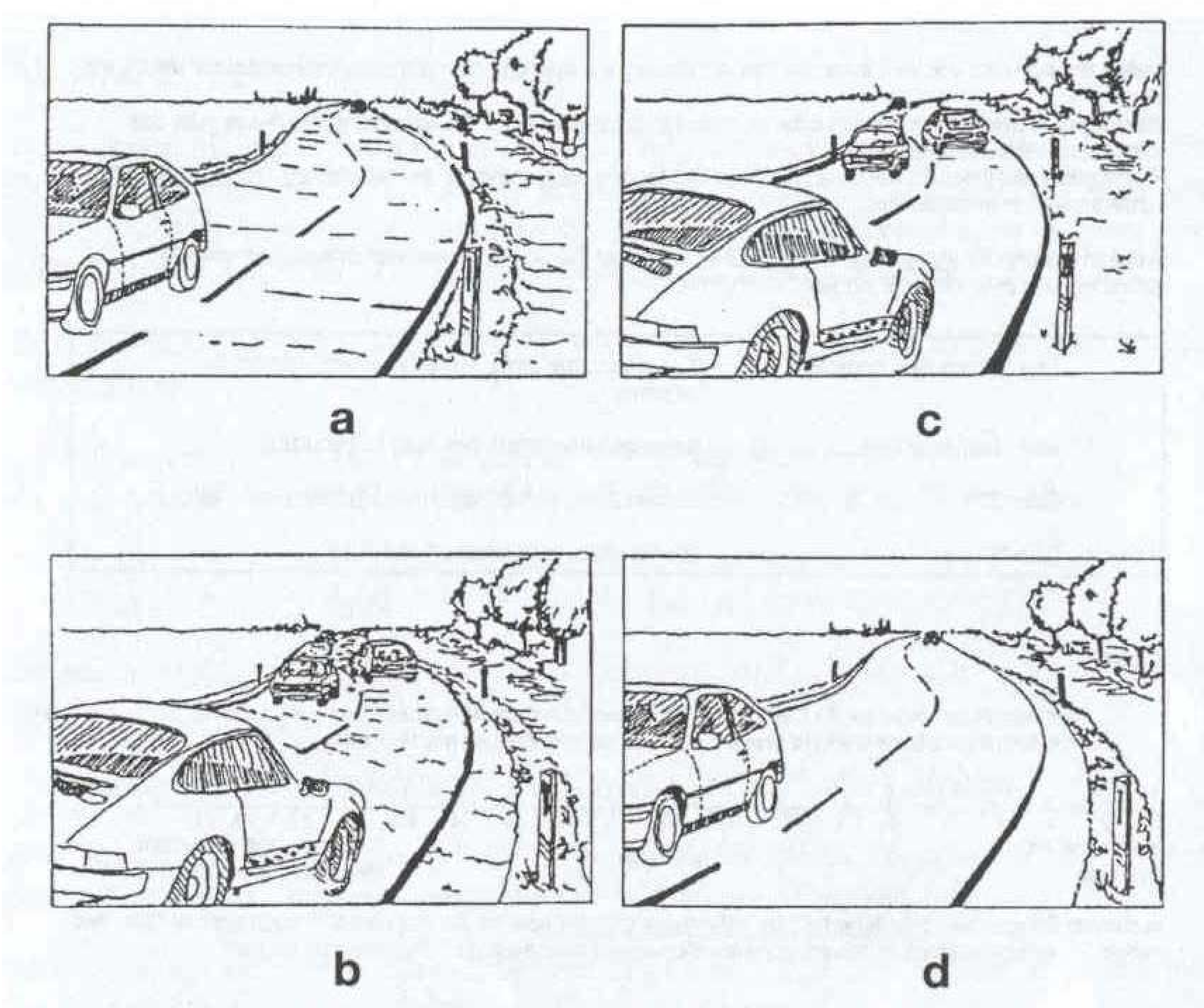
Gegeben ist folgende Situation:		
Öffentlicher Verkehr:	Es fährt eine	Strassenbahn
	Strassenbahn fährt	alle 6 min
	Strassenbahn ist	in 0 von 10 Fällen unpünktlich
	Umsteigen	nein
	Fahrt dauert	insgesamt 20 min
	Fusswege von/zur Haltestelle dauern	insgesamt 7 min
	Fahrt mit der Strassenbahn kostet	2.50 DM
Rad:	Fussweg bis zum Rad	1 min
	Fahrzeit mit dem Rad ist	8 min
	Zum abstellen des Rades gibt es	keinen Fahrradständer
	Fussweg vom abgestellten Rad zum Ziel ..	1 min
	Als Radweg ausgebaut sind	15 % der Strecke
zu Fuss:	Gehzeit ist	23 min

Ihre Entscheidung wäre: Strassenbahn ___ Rad ___ zu Fuss ___

Example: *stated ranking*



Example: *stated ranking*



Modelling choices

Aggregate:

- Any suitable regression model
- Any suitable time-series model

Disaggregate:

- Random utility models
- Discriminant analysis
- SEM
- Classification trees
- Rule-based systems

General model

Separation of

- Knowledge and relevance of alternatives
- Choice among relevant alternatives

$$P_q(i) = P_q(i \mid C)\, P_q(C \mid C^*)$$

bei

$P_q(j)$	Choice probability of alternative j by person q
$P_q(j \mid C)$	Choice probability of alternative j by person q from the set of alternatives C
$P_q(C \mid C^*)$	Choice probability of C by person q from the set of all sets of alternatives C*

Definition of the utility

We assume that the utility U_{jq} of alternative j for person q:

$$U_{jq} = U(X_{kjq}) = \eta \, V(X_{kjq}) + \varepsilon_{jq}$$

$V(X_{kjq})$	Systematic part measurable for the observer
ε_{jq}	Non systematic part, which is random for the (outside) observer
η	Scale parameter of the distribution of ε_{jq} $\eta = 1$ for identification

Stochastistic utility maximisation

We assume:

$$U_{jq} \geq U_{iq} \;, \; \forall i$$

$$V_{jq} - V_{iq} \geq \varepsilon_{iq} - \varepsilon_{jq} \;, \; \forall i$$

from which follows:

$$P_{jq} = Prob\{ \varepsilon_{iq} \leq \varepsilon_{jq} + (V_{jq} - V_{iq}) \} \;, \; \forall i$$

$$P_{jq} = \int f(\varepsilon)\, \mathrm{d}\varepsilon$$

Formulation of the utility function

We assume

$$
\begin{aligned}
V(X_{kjq}) = \; & a_j + \\
& ß_{j,q,t} * x_{j,q,t} + \\
& ß_{q,t} * c_{q,t} * x_{j,q,t} + \\
& ß_q * p_q * x_{j,q,t}
\end{aligned}
$$

or

$$
\begin{aligned}
V(X_{kjq}) = \; & a'_j + a_{q,t} * c_{q,t} + a_q * p_q \\
& ß_{j,q,t} * x_{j,q,t} + \\
& ß_{q,t} * c_{q,t} * x_{j,q,t} + \\
& ß_q * p_q * x_{j,q,t}
\end{aligned}
$$

Generalised Extreme Value - models

Generator of the Logit – family of models (McFadden, 1978):

$G[y_1, y_2, \ldots y_n]$ has for $y_n \geq 0$ the following properties:

- G is non-negative
- G is homogenous to the degree ϕ; i.e.
  $G[\alpha\, y_1, \alpha\, y_2, \ldots \alpha y_n] = \alpha^{\phi}\, G[y_1, y_2, \ldots y_n]$
- $\lim_{(y_i \to \infty)} G[y_1, y_2, \ldots y_n] = 0$; for all i
- The l^{th} derivative of G with respect to an arbitrary combination of l y_i´s is non-negative, if l is even and non-positive, if l is odd.

Generalised Extreme Value - Model

The following stochastic model is then consistent with the principle of utility maximisation

$$P_q(j) = \frac{e(V_{jq})\ G_i[e(V_{1q}), e(V_{2q}), \ldots\ e(V_{nq})]}{\phi\ G[e(V_{1q}), e(V_{2q}), \ldots\ e(V_{nq})]}$$

$G_i[y_1, y_2, \ldots\ y_n]$ is the first derivate of G with respect to y_i

Assumptions needed and degrees of freedom

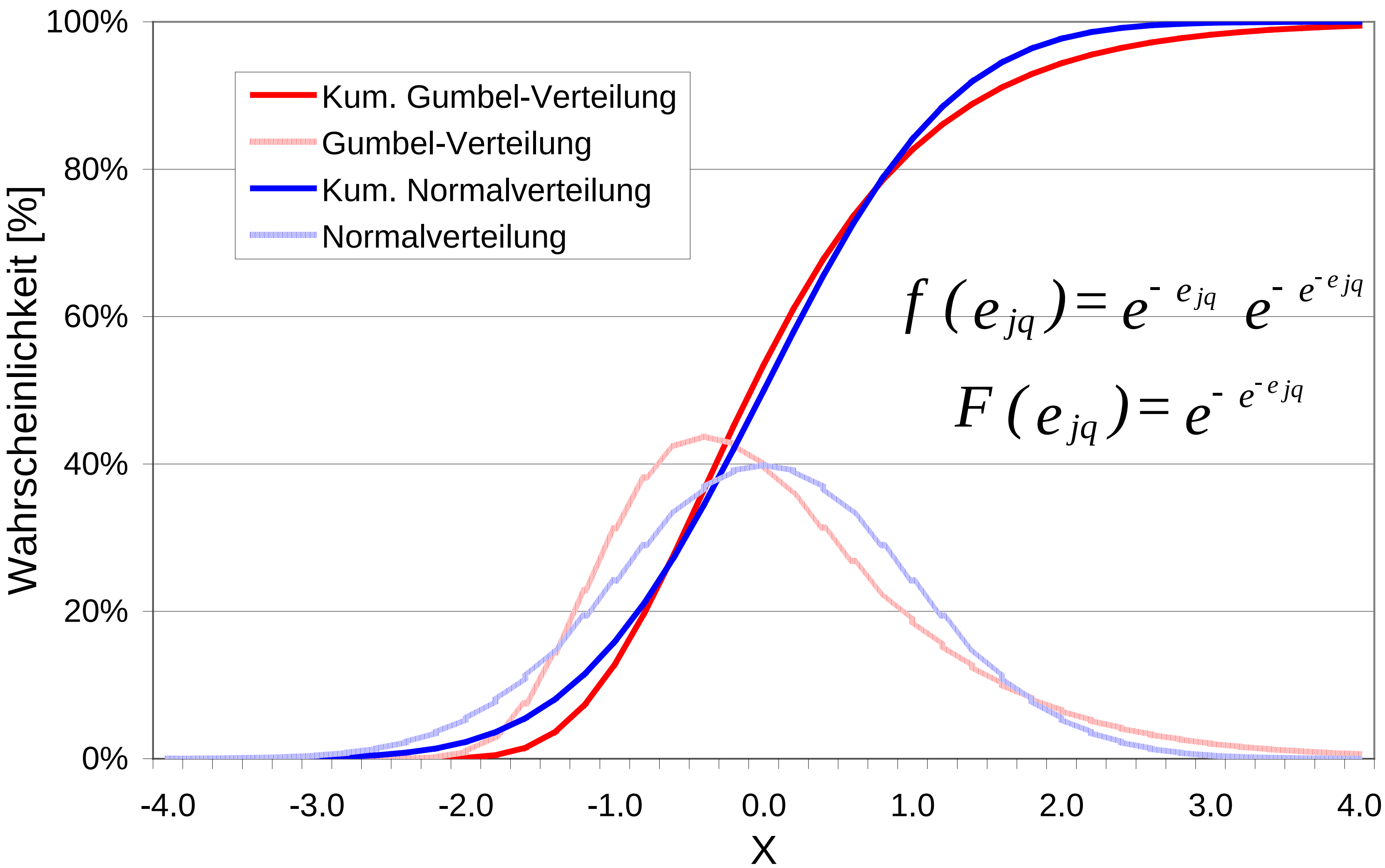
The following a-priori assumptions are needed

- Specification of V_{jq}
- Distribution of ε_{jq}
- Form of G[]

Degrees of freedom:

- Variance-covariance-matrix of ε_{jq} , respectively the patterns of similarity
- Distribution of β_{kjq}
- Functional form of the x_{kjq}

Logit assumption ε_{jq}: cumulative Gumbel-distribution



Parameter of the Gumbel-distribution

Fixing

Mode $= \mu$
Scale $= \eta$

one obtains

$$e_{jq} = \mu\,(\varepsilon_{jq} - \eta)$$

Mean $= \eta + 0.577/\eta$
Variance $= \pi^2/\,6\eta^2$

Structure of the variance-covariance-matrix Ω

How similar are the ε_{jq} ?

$$\Omega_{ij} = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

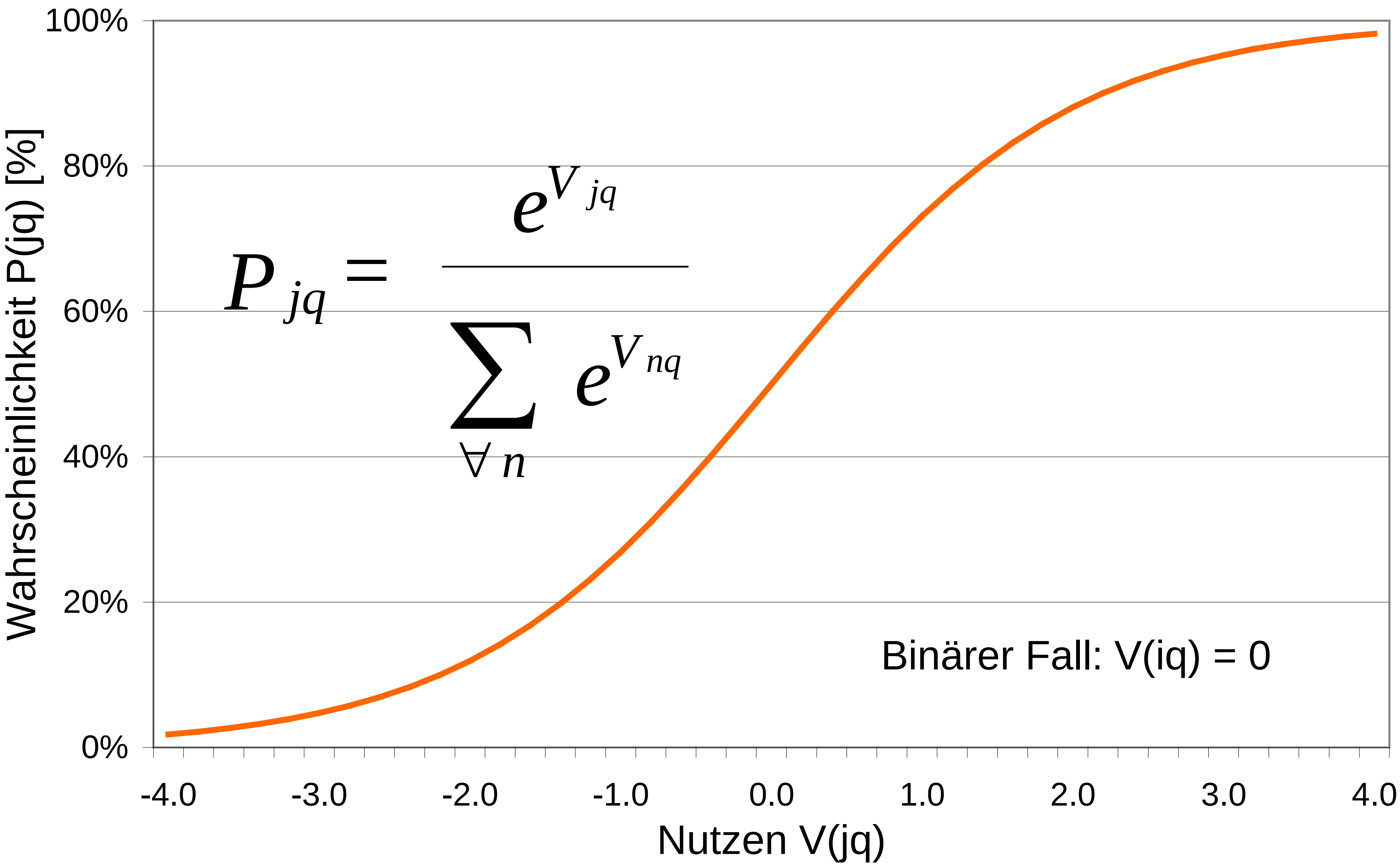
$$\Omega_{ij} = \begin{bmatrix} \sigma^2_{11} & -\sigma^2_{21} & \cdots & -\sigma^2_{n1} \\ \sigma^2_{21} & \sigma^2_{22} & \cdots & -\sigma^2_{n2} \\ \vdots & & & \\ \sigma^2_{n1} & \sigma^2_{n2} & \cdots & \sigma^2_{nn} \end{bmatrix}$$

Constant and independent or not

Example: Multinomial Logit (MNL) - model

- $G = \sum_{(\forall j)} y_j^{\mu}$ and $\mu = 1$
- ε_{jq} are Gumbel-distributed with $\Omega_{ij} = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & 1 \end{bmatrix}$
- β_{kjq} are constant across persons q and alternatives j
- V() is the result of the specification search

Example: binary MNL - model



$$P_{jq} = \frac{e^{V_{jq}}}{\sum_{\forall n} e^{V_{nq}}}$$

MNL: *Irrelevance of independent alternatives*

IIA: the ratio of the choice probabilities of any two alternatives is independent of the total number of alternatives

$$
\begin{aligned}
\frac{P_i}{P_j} &= \frac{e^{V_i}}{\sum_{\forall n} e^{V_n}} \frac{\sum_{\forall n} e^{V_n}}{e^{V_j}} \\
&= e^{V_i - V_j} \\
&= Const
\end{aligned}
$$

Capturing similarity through structure of Ω

A-priori assumptions about the structure of Ω

- Nested Logit (NL) (Alternative -> Nest)

- Cross-Nested-Logit (CNL) (Alternative -> Nests)

- Ordered Logit (OL) (Alternative -> Nested according to some natural ranking/order)

- Generalised Nested Logit (GNL)
- Network GEV - Model (NGEV)

Error components approach

We can add

$$U_{jq} = U(X_{kjq}) = \eta\ V(X_{kjq}) + \phi_{jq} + \varepsilon_{jq}$$

With

ϕ_{jq} ~ Normal distributed (using some factor structure)

ε_{jq} ~ Gumbel distributed as before

Assumptions about β_{kjq}

Is

$\beta_{kjq} = \beta_{kj*}$ = constant across all persons q ?

$\beta_{kjq} = \beta_{k**}$ = constant across persons q and alternatives j ?

or

β_{kjq} ~ a-priori set distribution with a mean β_{kj*} and Variance $\sigma^2_{\beta_{kj*}}$ over persons q alternative j and variable k

or is

$\beta_{kj*} = f(\sum X_{kjq})$, a function of other variables

Mixed and error component logit

Assuming distributed parameters results in

- Mixed MNL (MMNL) or random parameter logit (RPL)
- Error components logit (ECL)

Other GEV models can be substituted for the MNL

Estimation

Maximum likelihood approaches:

- GEV family using analytical approaches
- MMNL and ECL using simulated maximum likelihood approaches

Implemented in SAS, (SPSS), LIMDEP, ALOGIT, BIOGEME, or various Gauss - codes

Swiss Value of Travel Time Savings (VTTS) study

Steering group organised by SVI (chair: U. Weidmann, SBB)

Core team:

- G Abay
- KW Axhausen
- A König

External advisers:

- JJ Bates
- M Bierlaire

Survey approach

Add-on to ongoing RP – survey (KEP of SBB)

Pretests:

- Route choice
- Mode choice
- Destination choice

Main study

- Route choice
- Mode choice

Route choice car (Main study)

Route A	Route B
Reisekosten: 18 Fr.	Reisekosten: 23 Fr.
Gesamtfahrzeit: 40 Min.	Gesamtfahrzeit: 20 Min.
davon in stop and go: 10 Min.	davon in stop and go: 5 Min.
davon freie Fahrt : 30 Min.	davon in freier Fahrt: 15 Min.

☐	← **Ihre Wahl** →	☐

Mode choice (Main study)

PW	Bahn
Reisekosten: 13 Fr.	Reisekosten: 23 Fr.
Gesamtfahrzeit: 30 Min.	Gesamtfahrzeit: 20 Min.
davon in stop and go: 5 Min.	Takt: 30 Min.
davon freie Fahrt : 25 Min.	Anzahl Umsteigen: 0-mal

☐ **← Ihre Wahl →** ☐

Response behaviour (Main study)

Interviewed for KEP:	5560	(during weeks 22 to 40 of 2002)
Willing to participate:	3216	(58% of interviewees)
Experments sent out:	2317	(72% of willing interviewees)
Returns:	1222	(53% of experiments sent out)

Response rate for the different experiments

Chosen mode	Car availability	RC chosen mode	Number of choice situations	Response rate
Car	Yes	Yes	15	52.2
Car	Yes	No	15	48.6
Bus	Yes	Yes	15	54.4
Rail	Yes	Yes	15	65.7
Bus	No	Yes	9	37.7
Rail	No	Yes	9	50.2

Sample drift: Shares by age, gender and education [%]

Variable	MZ 2000	KEP	Willing to participate	Returns
Females	51	54	50	41
Below 35	40	26	28	26
36-55	32	40	41	49
Above 55	28	34	31	26
Regular schooling	32	22	16	10
Professional training	41	53	54	46
Tertiary education	27	27	30	44

Sample drift: Shares by mobility tools and income [%]

Variable	MZ 2000	KEP	Willing to participate	Returns
Discount card	35	38	43	54
National season	6	6	7	11
Car available	77	63	62	73
Up to 40 kSFr	21	-	-	19
40 – 80 kSFr	42	-	-	35
80 – 125 kSFr	27	-	-	33
125 and more kSFr	11	-	-	14

Modelling strategy

Experimental variables only

+ Inertia indicators
+ Socio-demographic variables
+ Distance and income elasticities
+ RPL for cost and travel times
+ Interaction with trip purpose

For each experiment and then for pooled estimates

Specification of the elasticities

Non-linear elements of the utility function can be specified in Biogeme:

$$\beta_{Cost} * \left(\frac{Income}{Mean\ income} \right)^{\varepsilon_{Income}} * \left(\frac{Trip\ length}{Mean\ trip\ length} \right)^{\varepsilon_{Trip\ length}} * Cost$$

(Adopted from recent reanalysis of UK VTTS study)

Biogeme

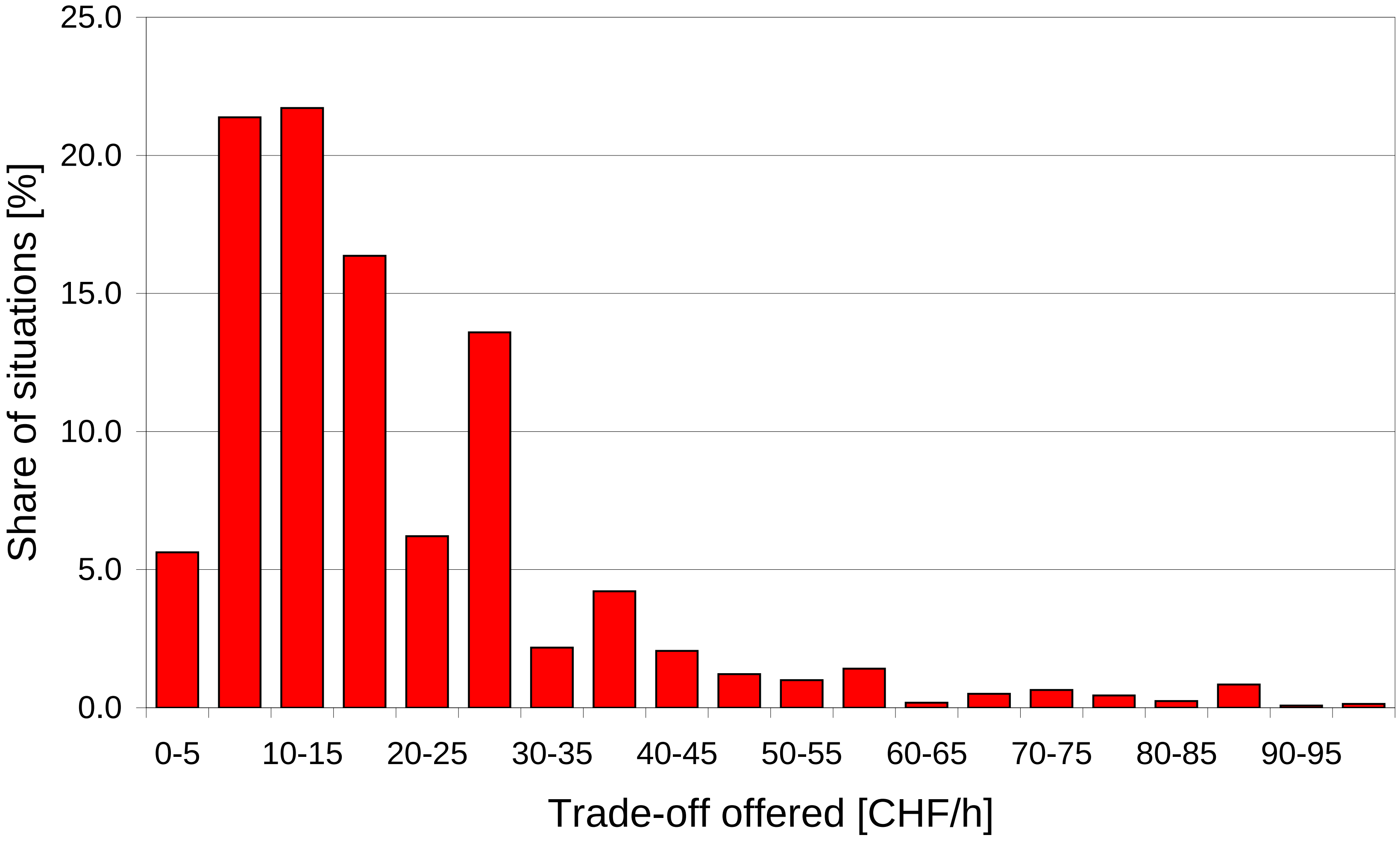
General GEV-modell estimation tool:

- MNL, NL, CNL, NetworkGEV
- Mixed logit
- Non-linear elements in the utility function, including Box-Cox transforms
- Direct estimation of error scales

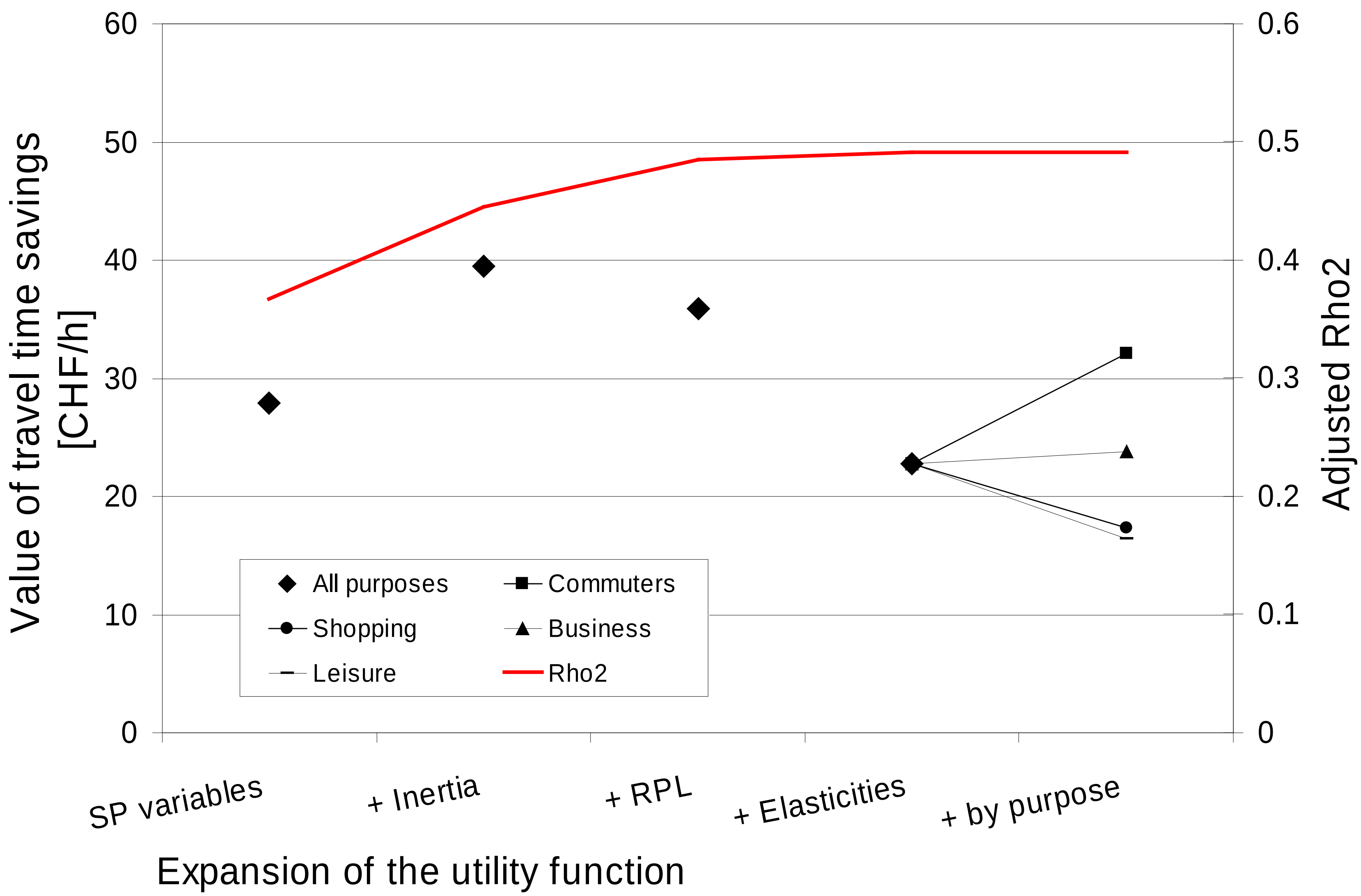
See for the freeware:

http://roso.epfl.ch/biogeme

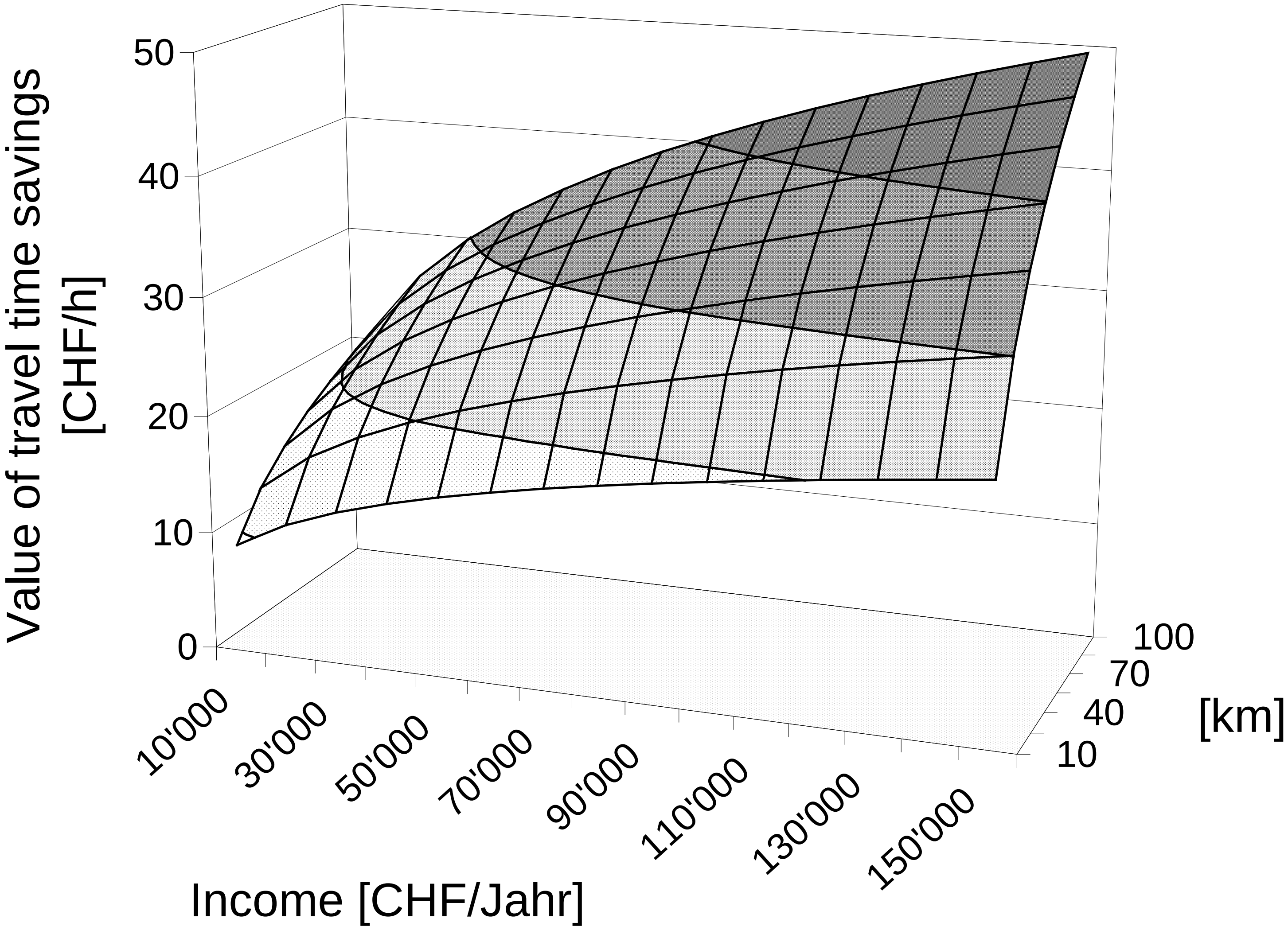
Trade-offs offered (Route choice rail)



Model trajectory: Mode choice – public transport user



Value of travel time savings: Car commuters



Fusing information

Problem: Multiple different views of the same parameters via different samples

Fusion while account for error scale differences, for example

$$U_{jq1} = U(X_{kjq1}) = \eta_1 V(X_{kjq1}) + \varepsilon_{jq1}$$
$$U_{jq2} = U(X_{kjq2}) = \eta_2 V(X_{kjq2}) + \varepsilon_{jq2}$$

with $\eta_1 = 1$

and one or more $\beta_{kjq1} = \beta_{kjq2}$

Relative scaling of the error distributions

Experiment	Route choice only		Combined	
	Para-meter	T-test	Para-meter	T-test
Mode choice	-	-	0.66	8.57
Route choice car	1.82	3.10	1.39	2.47
Route choice rail (car users)	0.97	1.19	1.05	0.83
Route choice rail (rail user)	1.00	-	1.00	-

Values of travel time savings at population mean

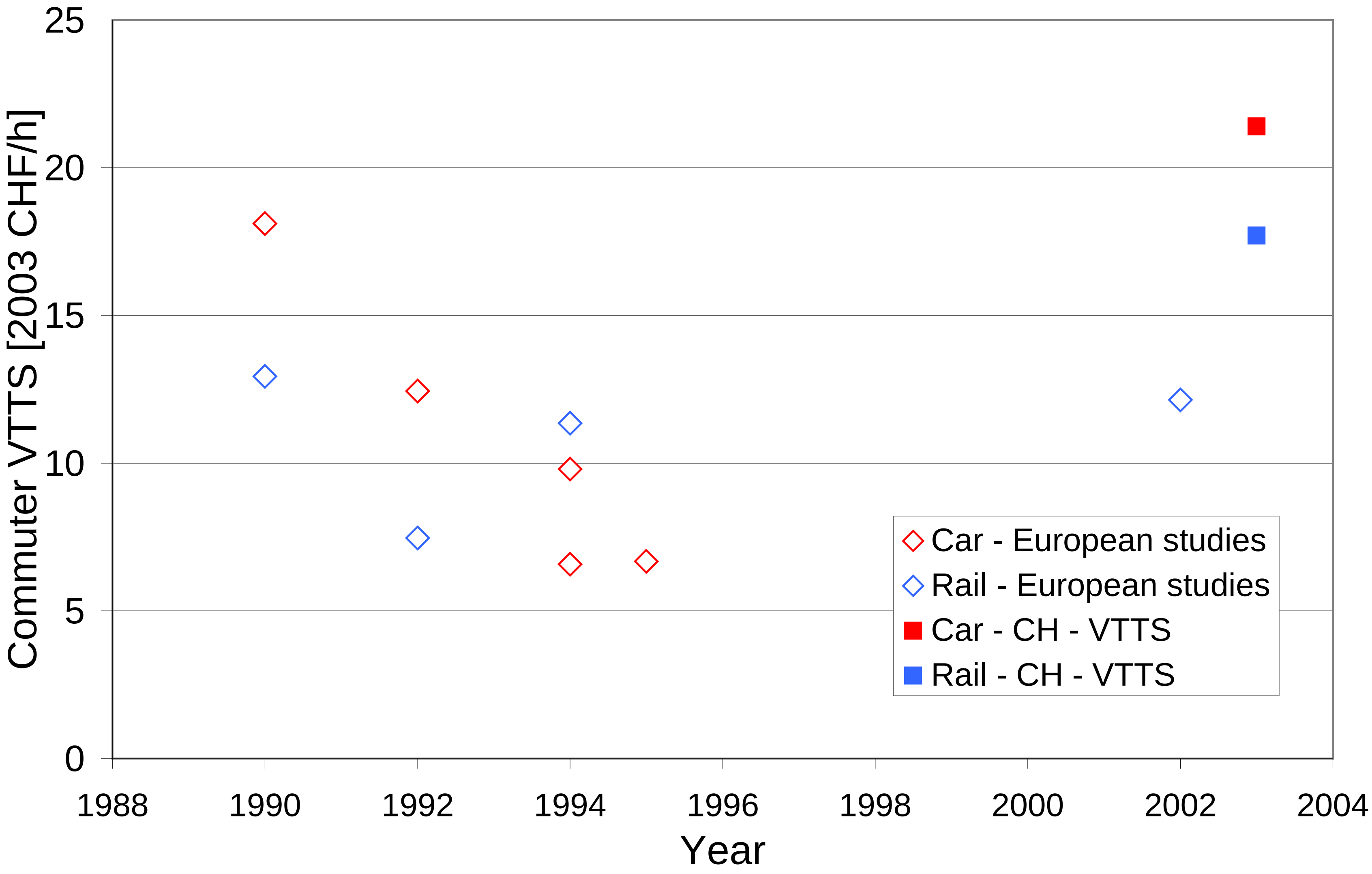
	Commuter		Shopping		Business		Leisure	
Car	21.4	2.9	18.1	3.8	32.5	-	12.3	0.8
Rail	17.7	1.7	13.8	2.1	30.3	-	9.7	0.5

Values and variances in [sFr/h];

Variances were computed by a Taylor expansion

VTTS for commuter is equivalent to 30-35% of average hourly wage

Comparison with other recent studies



Challenges and outlook: Data collection

RP:

- Expanding the variable set
- Maintaining response rates
- Incorporating tracing technologies

SP:

- Stability of results
- Size of experiment (number of variables, number of choice situations)
- Adaptive designs
- Hierarchical designs

Challenges and outlook: Choice modelling

- Measuring similarity in V()
- Identification of the structure of Ω from the observations
- Identification of the choice sets

- Non-parametric MMNL and suitable choice of parametric distributions
- Speed of estimation

- Stability of substantive results

- Integration into agent-based simulations

Literature

Axhausen, K.W., A. König, G. Abay, J.J. Bates and M. Bierlaire (2004) Swiss value of travel time savings, paper presented at the 2004 European Transport Conference, Strasbourg, October 2004

Axhausen, K. W. und G. Sammer (2001) Stated responses: Überblick, Grenzen, Möglichkeiten, *Internationales Verkehrswesen*, **53** (6) 274-278**.**

Ben Akiva, M.E. and S.R. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge.

Bhat, C. R. (2001) Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model, *Transportation Research,* **35B** (7) 677-693.

König, A. and K.W. Axhausen (2004) Zeitkostenansätze im Personenverkehr, final report for SVI 2001/534, *Schriftenreihe*, **1065**, Bundesamt für Strassen, UVEK, Bern.

Louviere, J.J. , D.A. Hensher and J.D. Swait (2000) *Stated Choice Methods: Analysis and Application*, Cambridge University Press, Cambridge.

Train, K. (2003) *Discrete Choice Analysis with Simulations*, Cambridge University Press, New York.