Preferred citation style

Axhausen, K.W. (2015) Data problems, modelling challenges, presentation at the Transport Studies Group, Tokyo Institute of Technology, June 2015.

.

Data problems, modelling challenges

KW Axhausen

IVT
ETH
Zürich

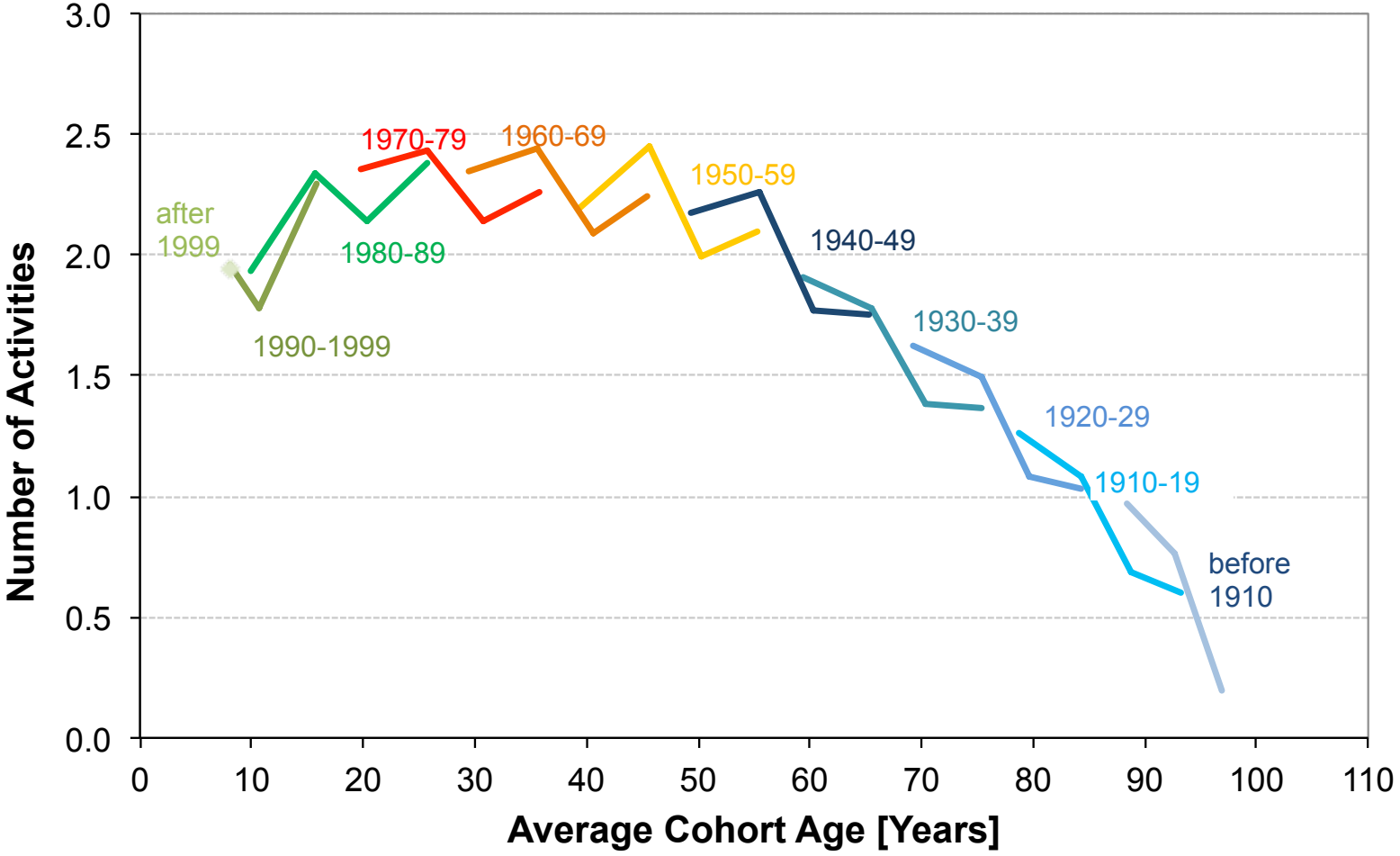
June 2015

Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

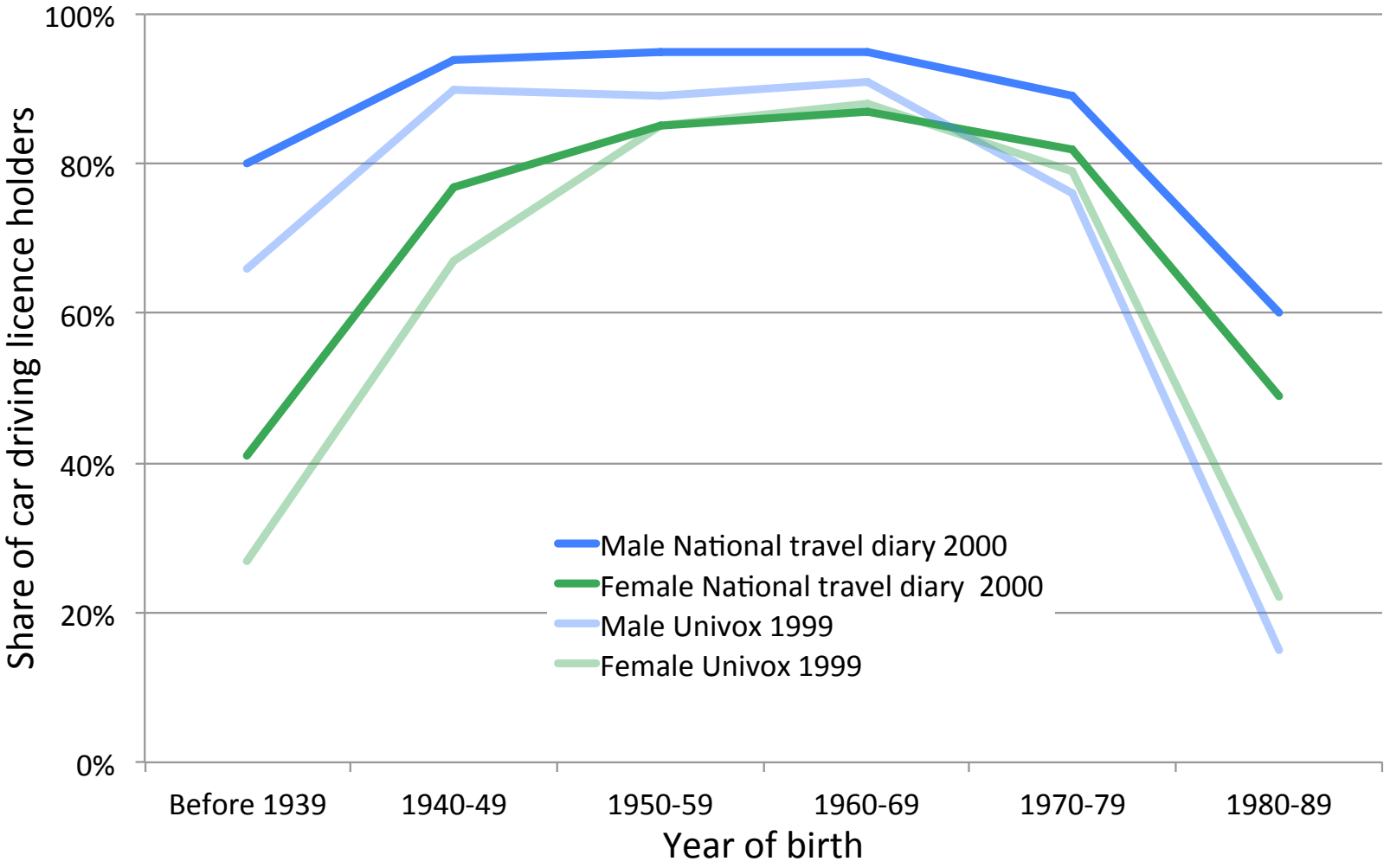
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Data challenges

Do we know the numbers? e.g. daily activities in Switzerland



Do we know the numbers? e.g. drivers licence ownership



Protocols and response

Surveys, observations are „talk“

Two speakers

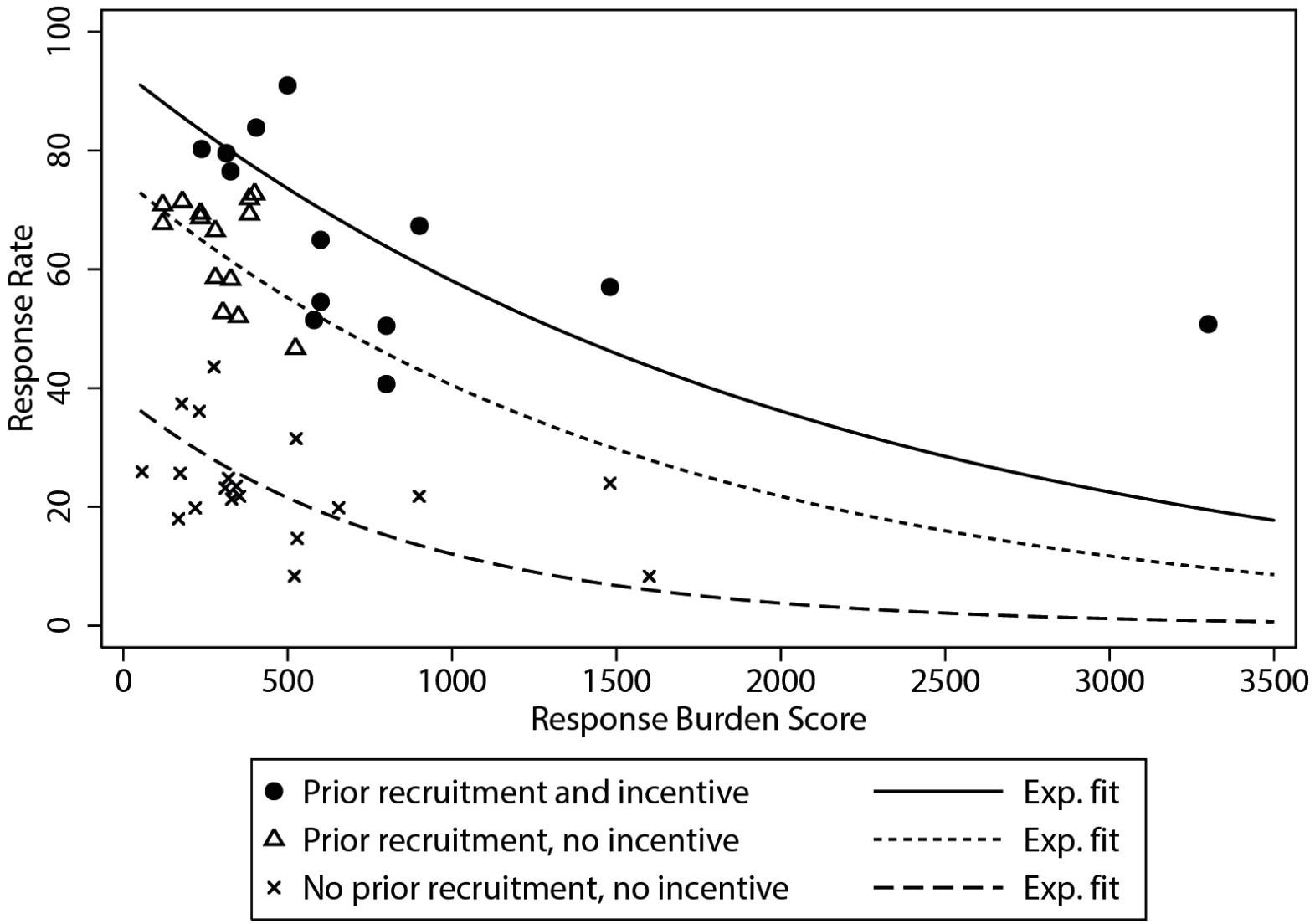
managing their „image“
staying within the rules of talking
staying within their socially allocated/identified role
fulfilling social expectations

talk and report with/to each other

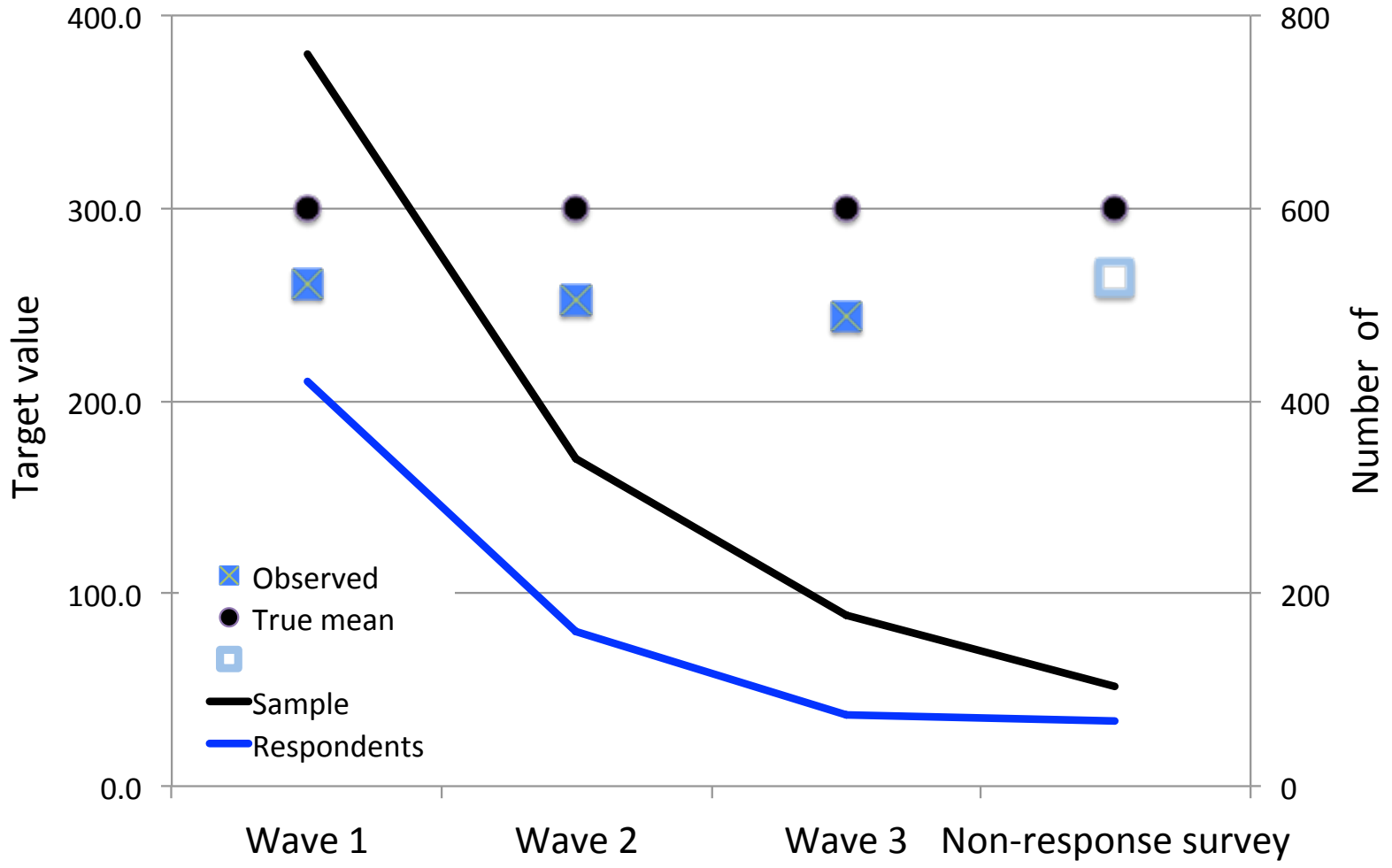
=>

„**Maintaing the willingness of the respondent to report**“

Response as a function of response burden @IVT, 2015

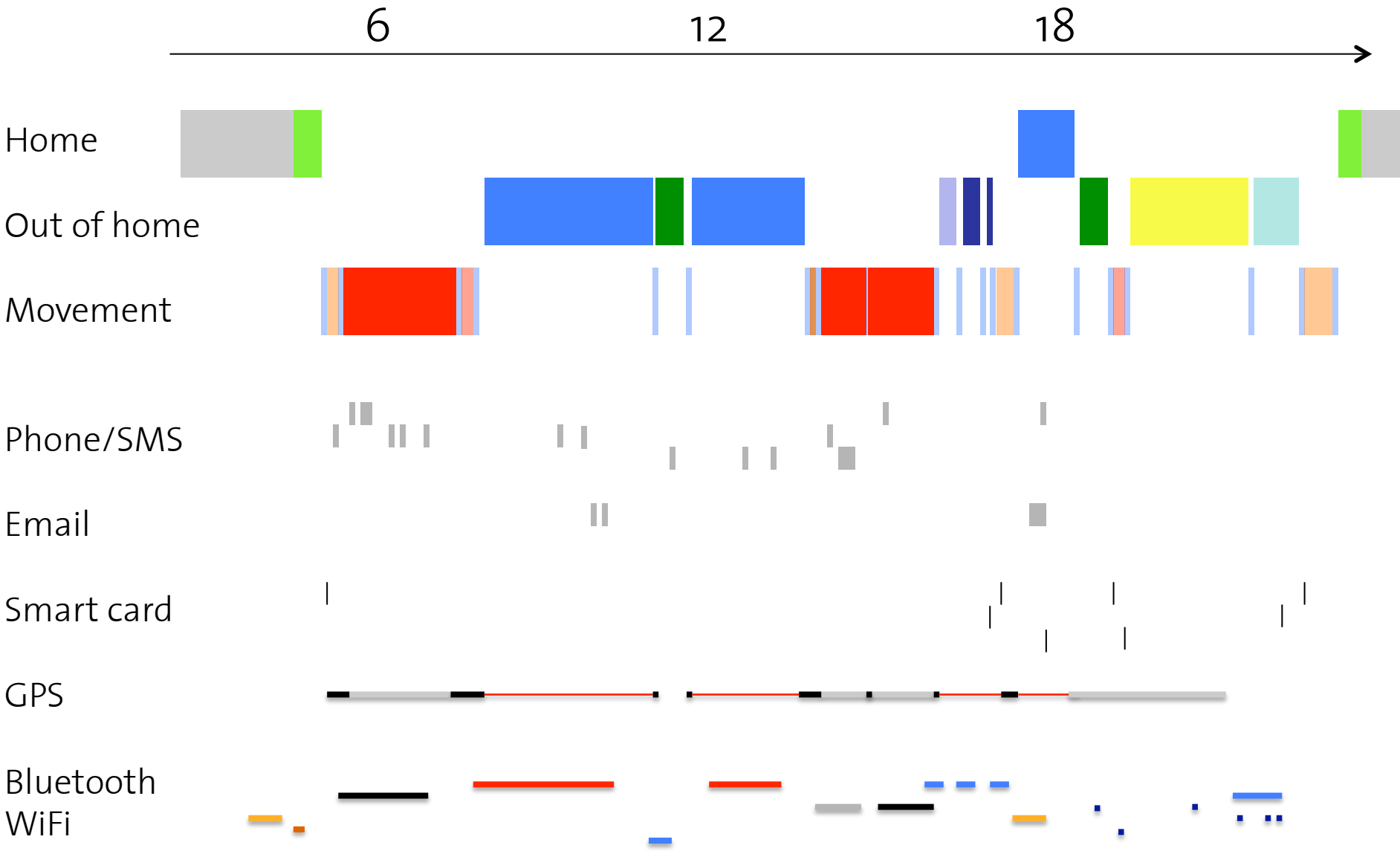


Response is a non-random process

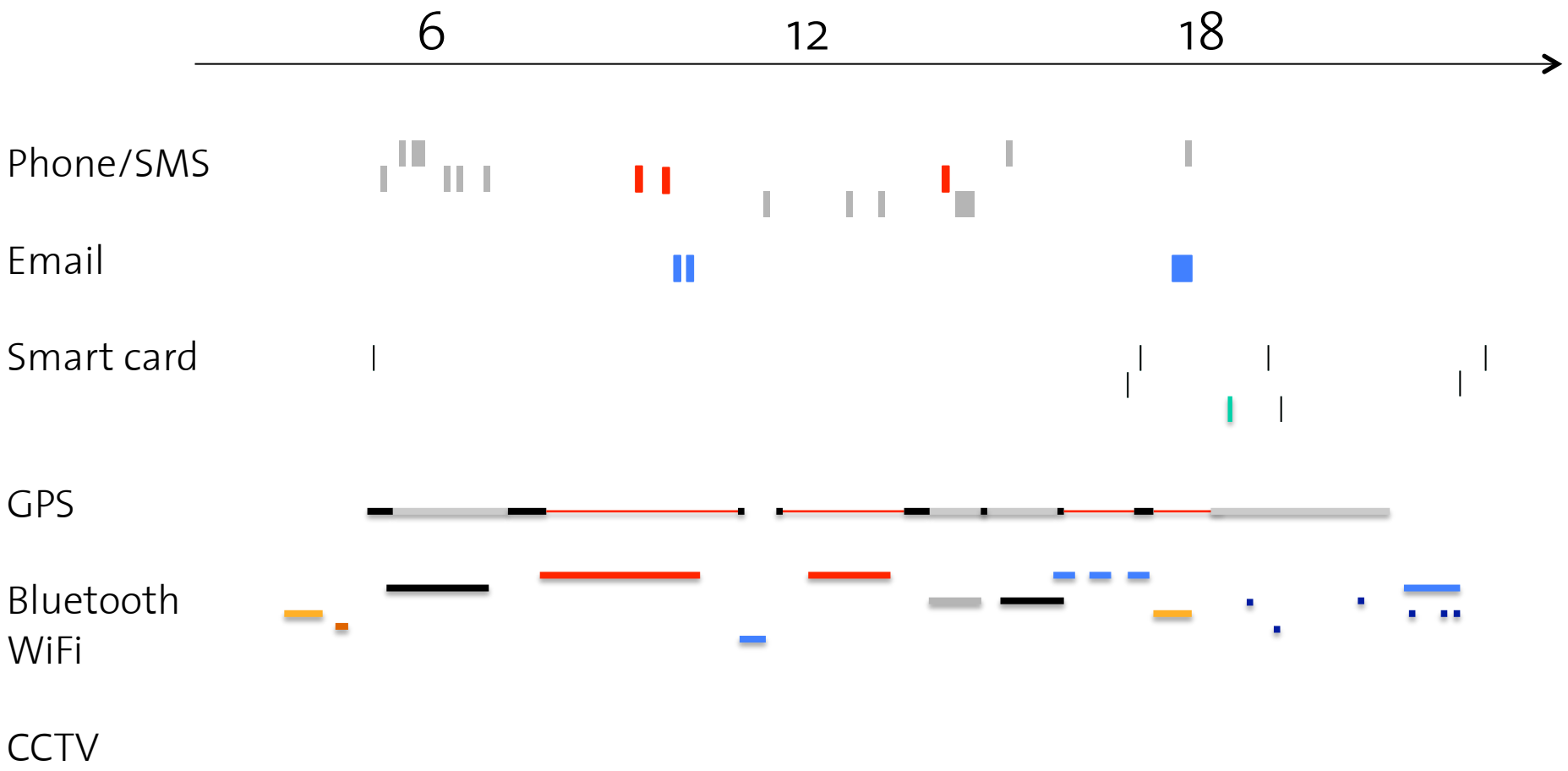


Known „error“ generating processes

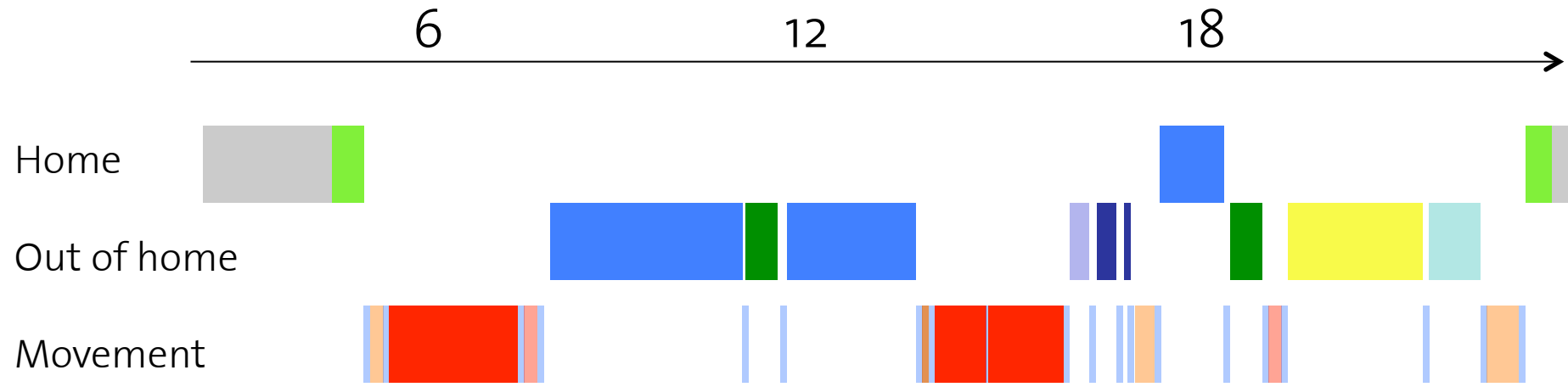
Activities, movement and traces: A full example record



Active/passive tracing: Many owners, locations, quality levels

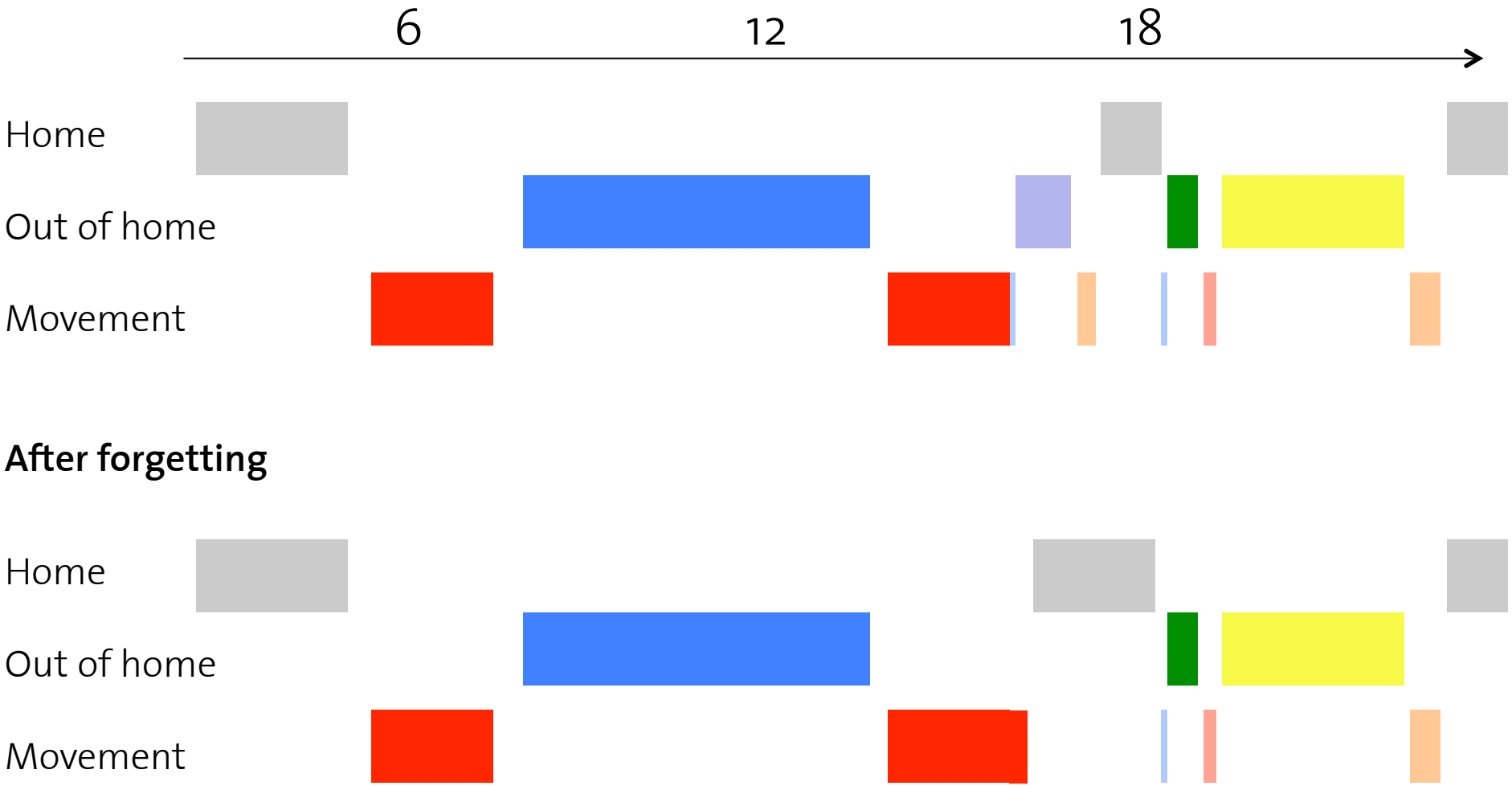


Filters imposed/suggested by the study: „Trips“

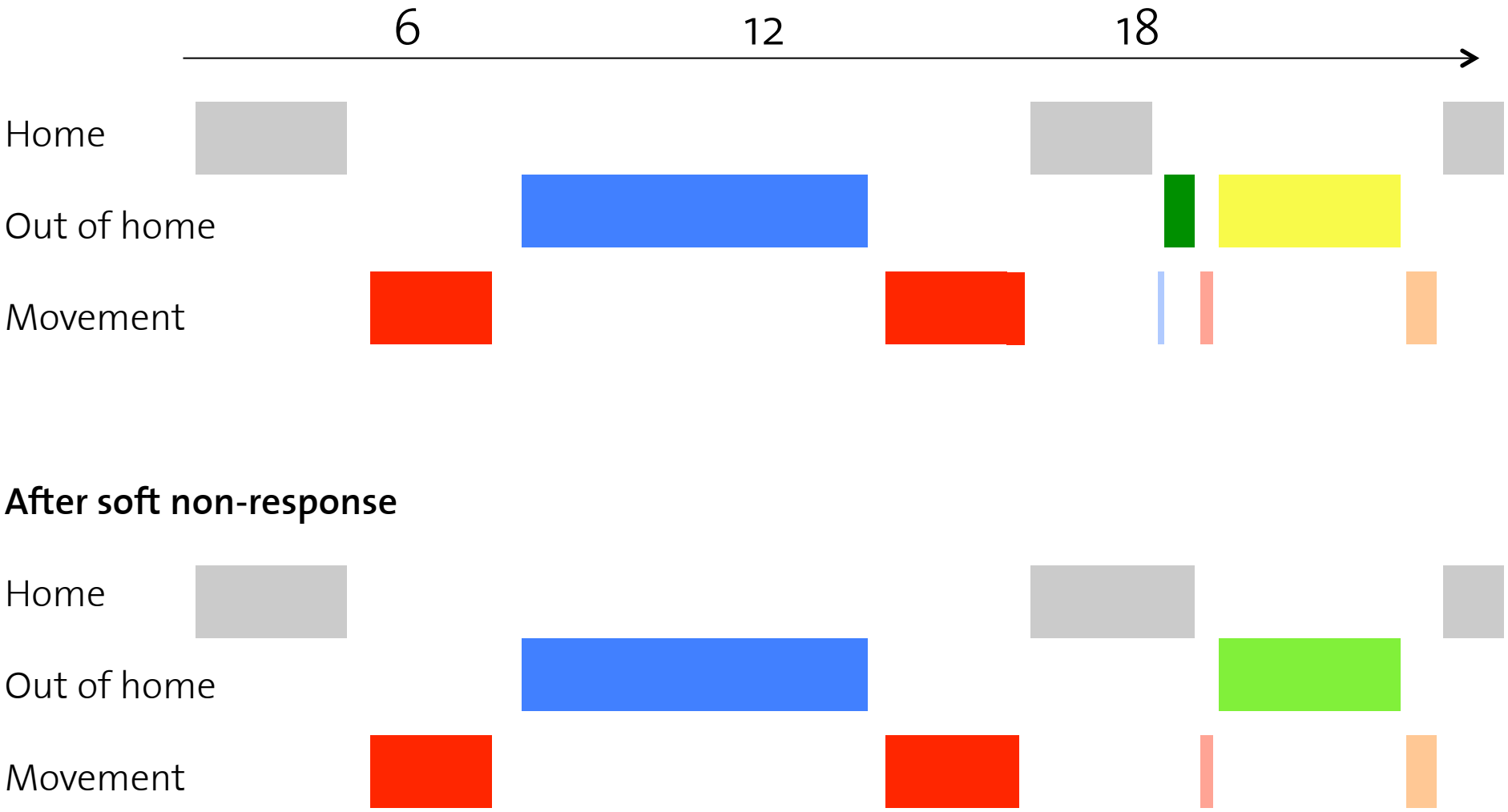


After “trip” filter:

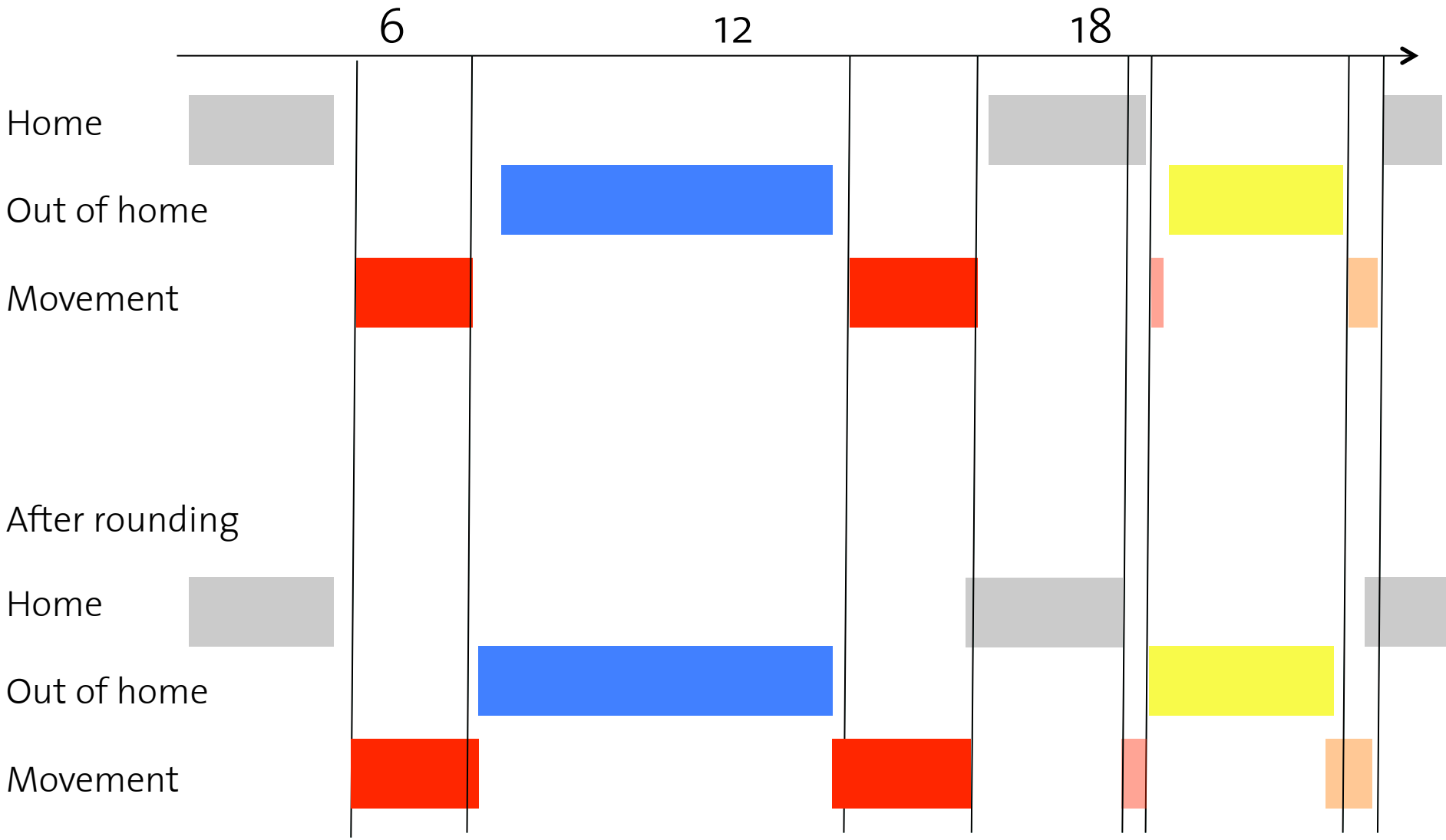
Filters due to the respondent: Forgetting



Filters imposed by the respondent: Soft non-response



Filters due to the respondent: Rounding



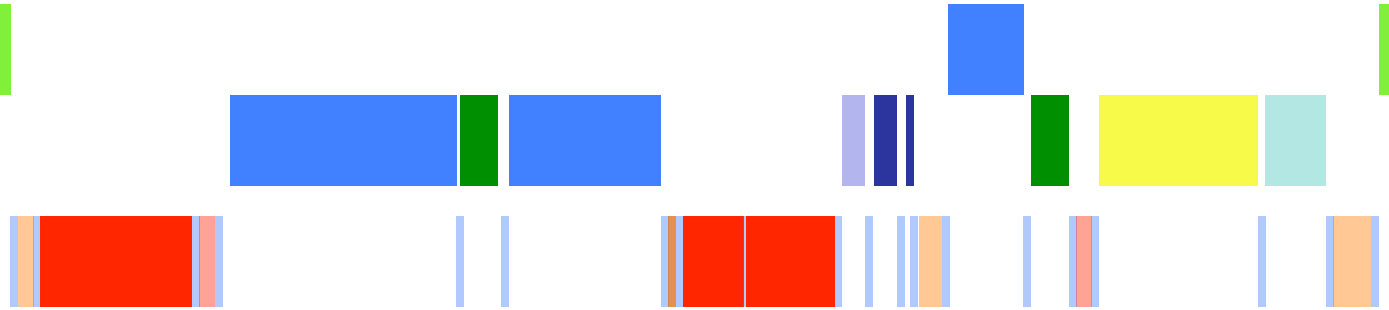
What is left ?

True

5 at home

9 Out of home

26 Stages,
11 trips,
1 subtour,
2 tours

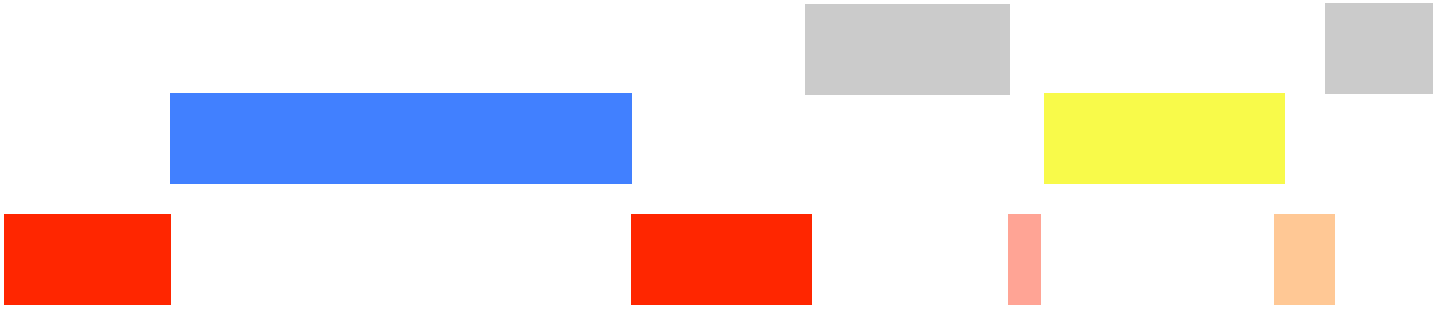


After all processes

3 at home

2 Out of home

4 trips,
2 tours



What happens next ?

Geocoding addresses

Ideal	Street addresses identifying the entry to the network
Best-case	Unambiguous street addresses
State of the art	Street address
State of practice	Street address/mid-street block/street corners; missing conversion of facility names
Still seen in practice	Arbitrary zonal centroid, e,g post offices

Calculating distances & travel time

Ideal	Complete GPS track for distance and times with pedestrian-networks added
Best-case	Minimal gaps, and state-of-the-art imputation of GPS tracks and modes
State of the art	SUE derived travel times and distances (navigation network)
State of practice	DUE derived travel times and distances (planning networks)
Still seen in practice	Shortest path on empty planning networks

What should we do ?

Next steps

- Query what we really need for
  - Cost-benefit analysis
  - Planning of prices and services
  - Planning for the slow modes
  - Social accounting

- High-quality multi-modal surveys to establish the measurement errors (add bluetooth and wifi senders, noise profile)
- Error correction models
- Cross check against third party sources

- Treat survey data as indicators in a measurement model
- Treat traces as indicators in a measurement model

, but especially

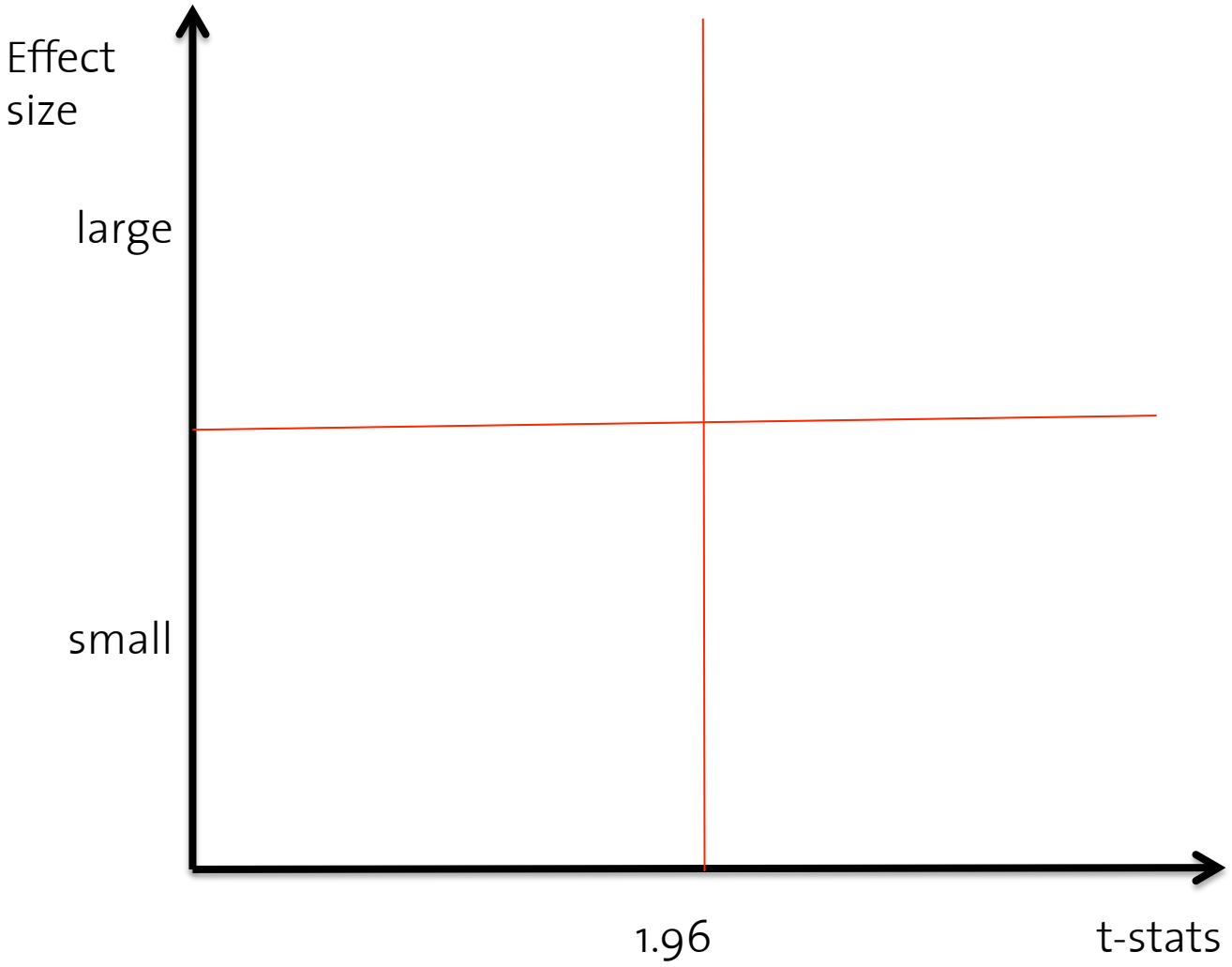
- Treat respondents as partners in a talk, discussion:
  - Frame your request in a way which addresses them in a clearly defined social role (citizen, driver, customer, etc.)
  - Account for their constraints (readability of text, full guidance through the forms, require no calculations – unless necessary, speak their 'language')
  - Be as complex, as the topic warrants, requires, but not more so
  - Don't surprise them with unannounced requests
  - Don't ask them to do work you can do
  - If appropriate, provide an incentive, acknowledgement

Modelling challenges

Modelling challenges: The usual worries

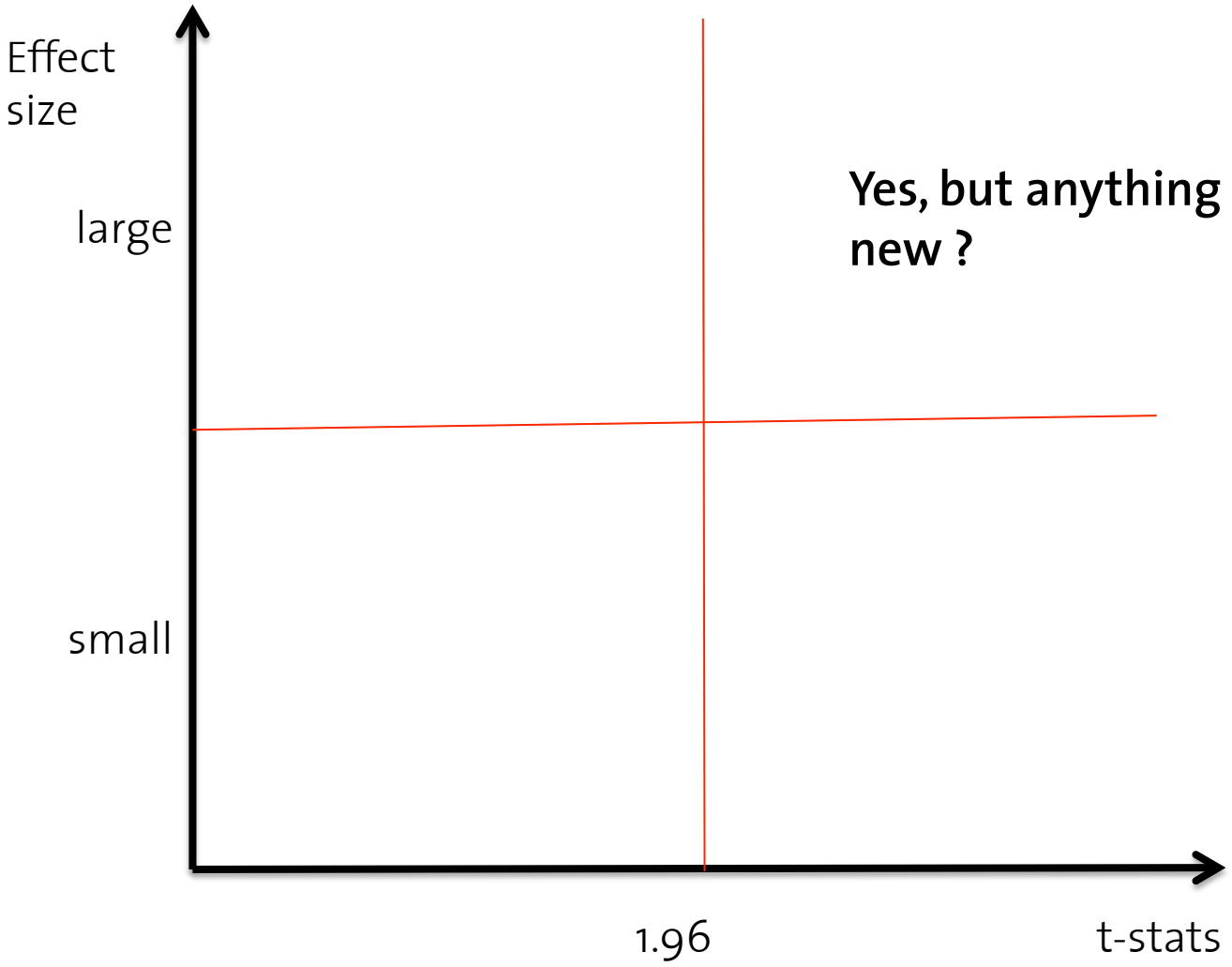
Error heterogenity	Is it always checked ?
Spatial correlations	Are they always checked ?
Temporal correlations	Are they always checked ?
Independence	Do we check the correlations of the independent variables (sample) thoroughly enough?
Endogenity	Do we fully account for it ? (sample selection)
Error of the second kind	Do you calculate it ?
Validation	How often do we ask for out-of-sample tests?
Substance	or do we talk about t-tests ?

Modelling challenges: Substance or t-tests ?



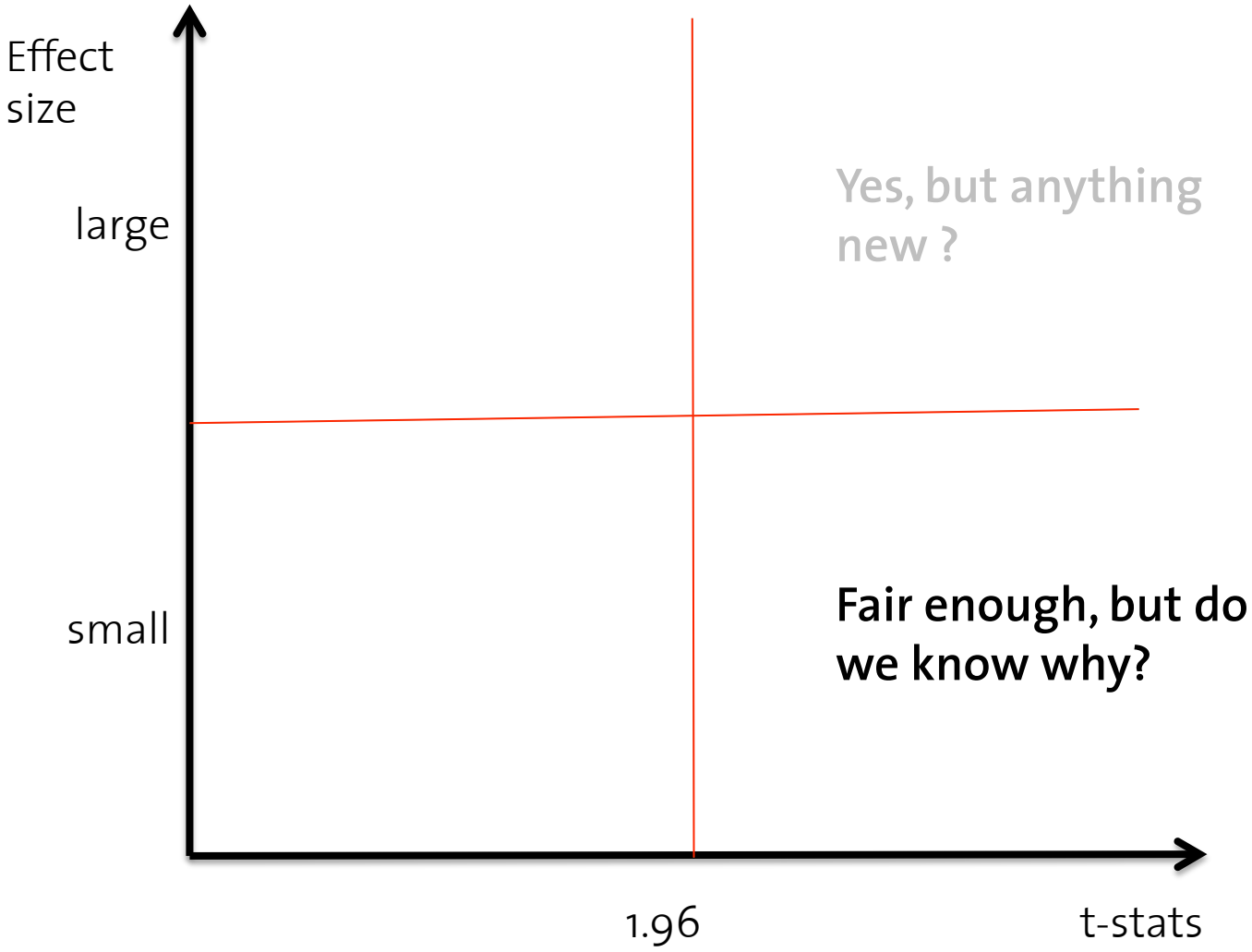
Adapted from Zilliak and McCloskey (2008)

Modelling challenges: Substance or t-tests ?



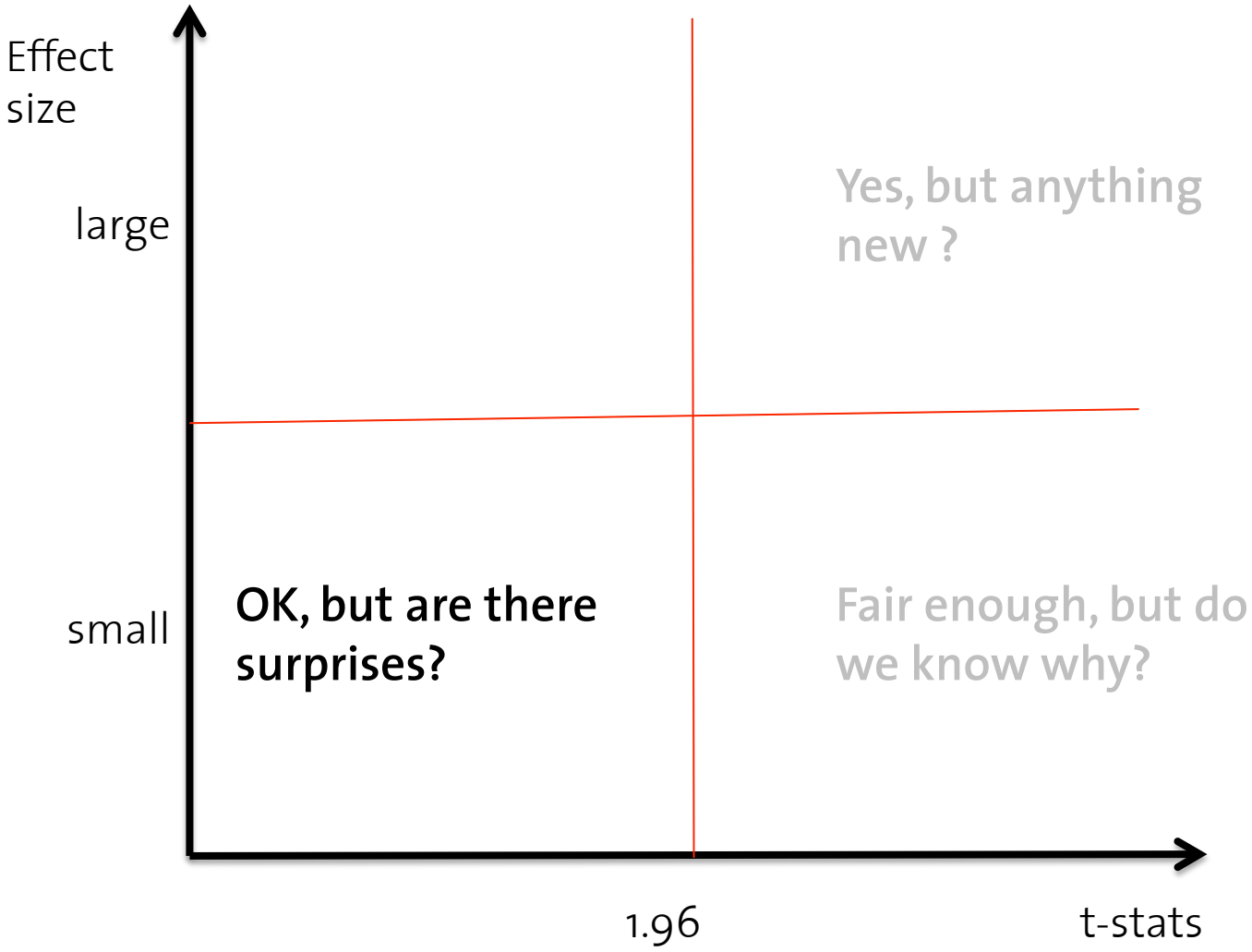
Adapted from Zilliak and McCloskey (2008)

Modelling challenges: Substance or t-tests ?



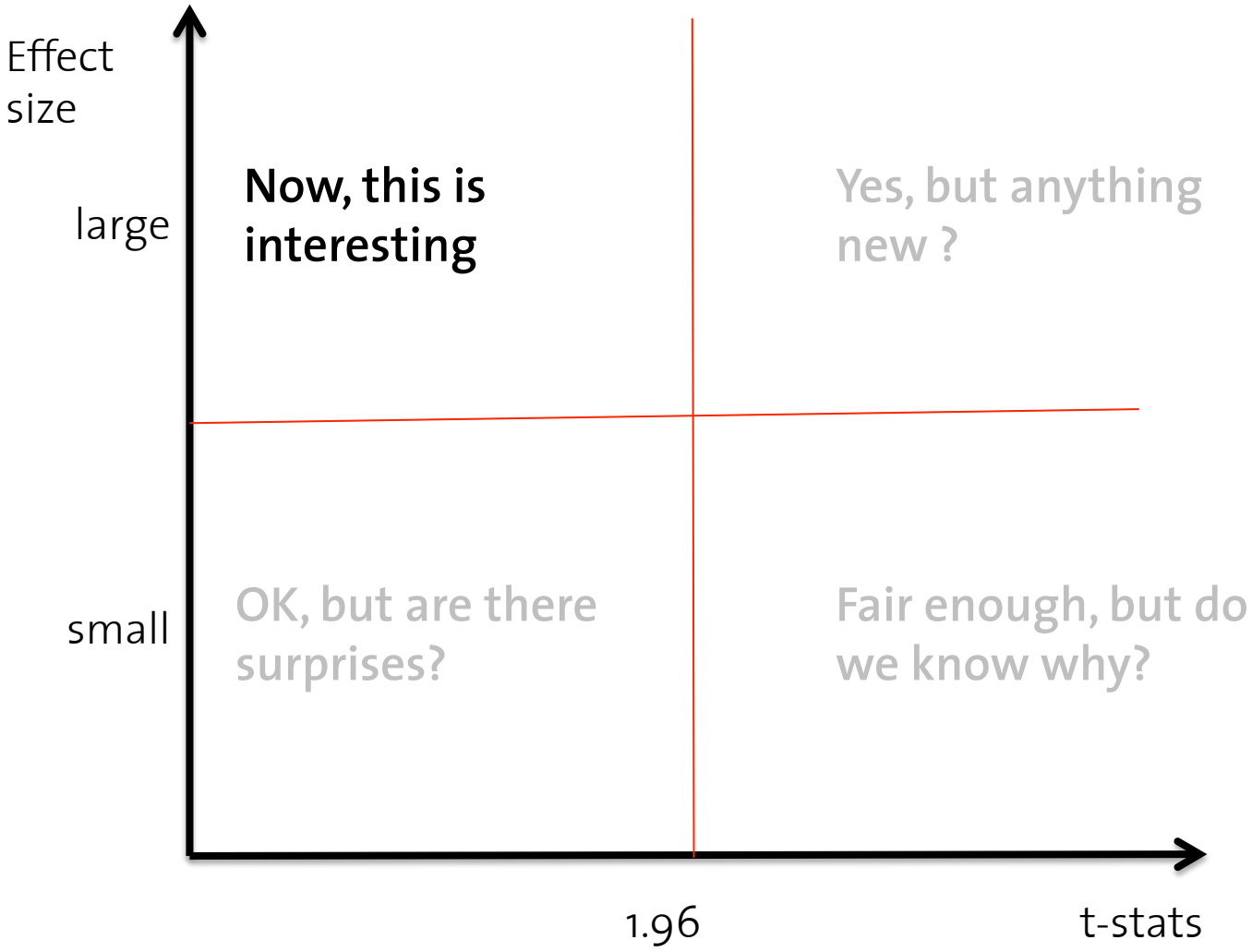
Adapted from Zilliak and McCloskey (2008)

Modelling challenges: Substance or t-tests ?



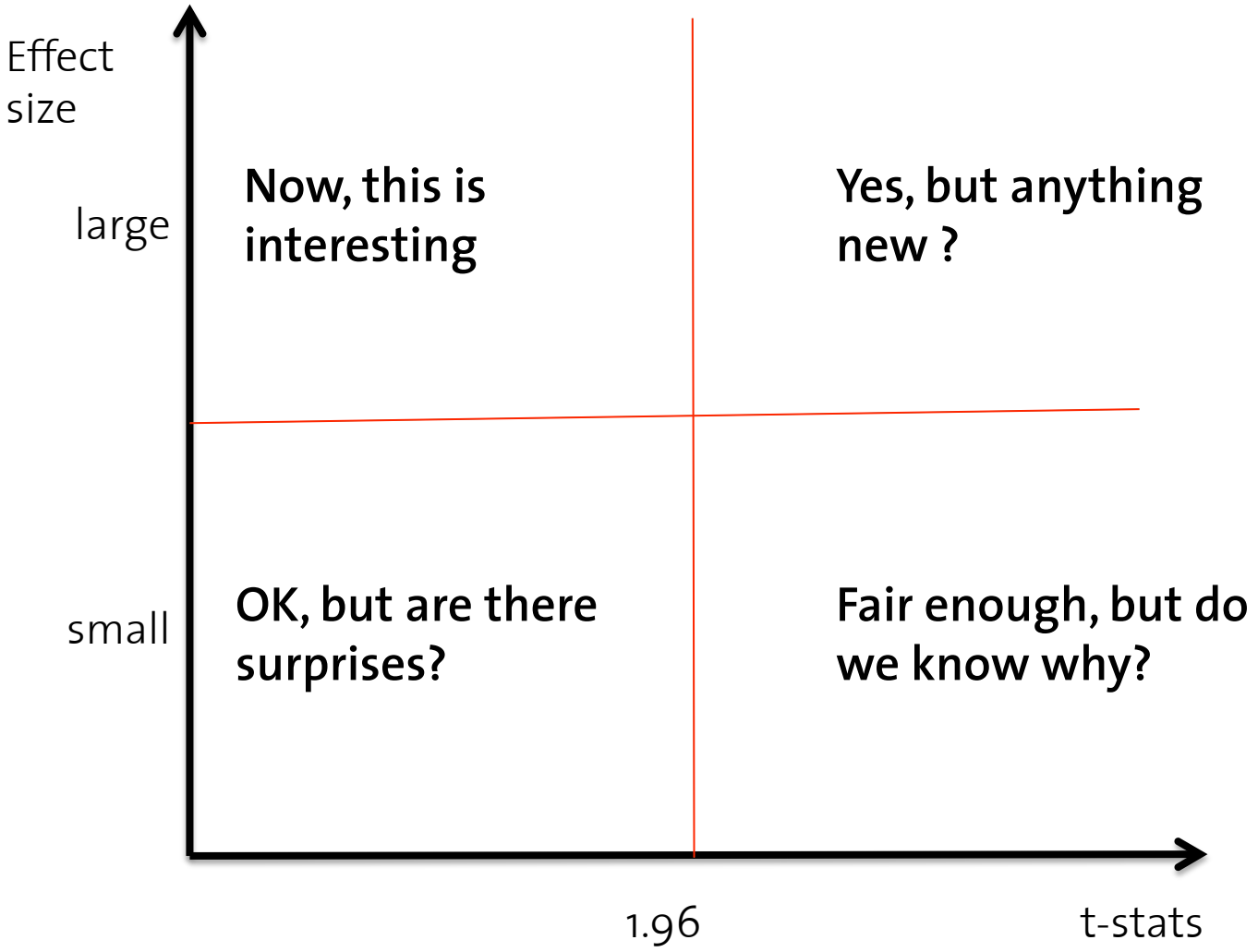
Adapted from Zilliak and McCloskey (2008)

Modelling challenges: Substance or t-tests ?



Adapted from Zilliak and McCloskey (2008)

Modelling challenges: Substance or t-tests ?



Adapted from Zilliak and McCloskey (2008)

Choice modelling challenges

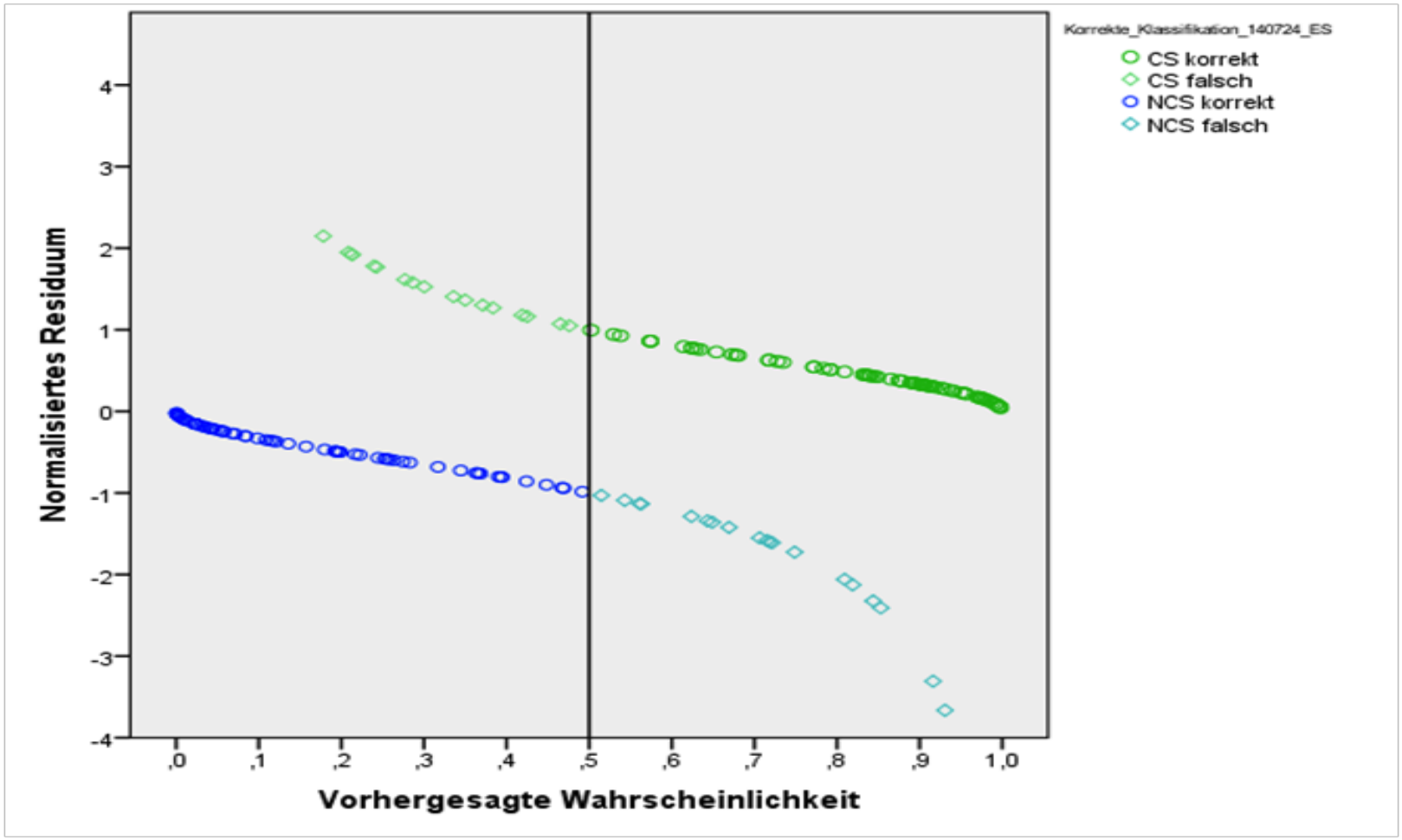
Choice modelling challenges: The usual worries

Error heterogenity	Is it always checked ?
Spatial correlations	Are they it always checked ?
Independence	Do we check the correlations of the independent variables (sample) thoroughly enough?
Endogenity	Do we fully account for it ? (sample selection)
Error of the second kind	Do you calculate it ?
Validation	How often do we ask for out-of-sample tests?
Substance	or do we talk about t-tests ?

Choice modelling challenges: less usual concerns

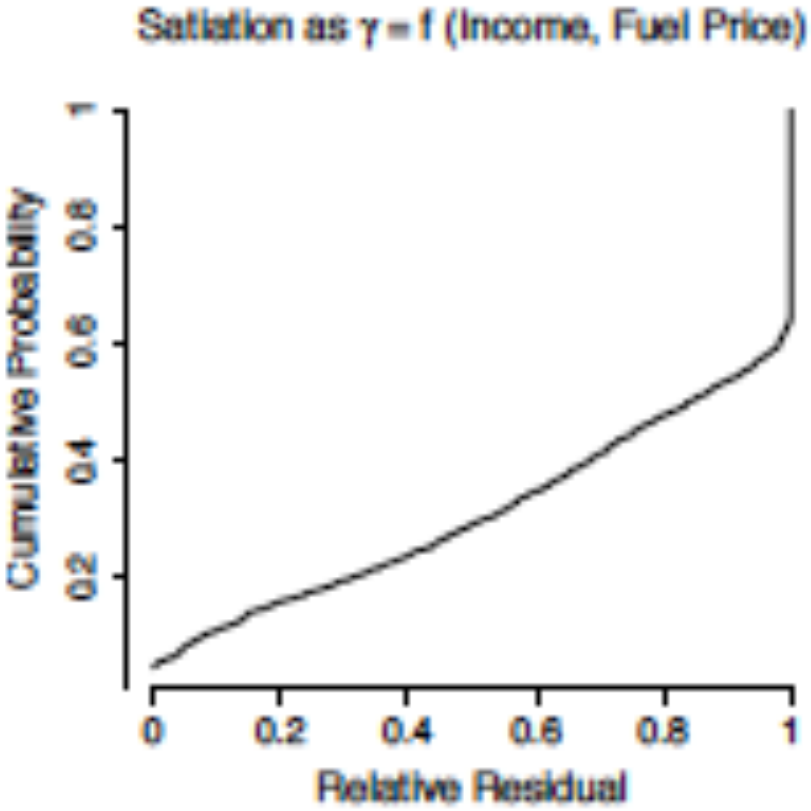
Error heterogenity	Why don't we check them ?
Number of non-chosen alternatives	How much leverage do they have for your problem?
Number of choice sets	How stable are our estimates?
Capacity constraints	Do we check for their impact on the parameters? (attribute values of the known (non)chosen alternatives)
Unit of analysis	Do we have a MAUP problem?

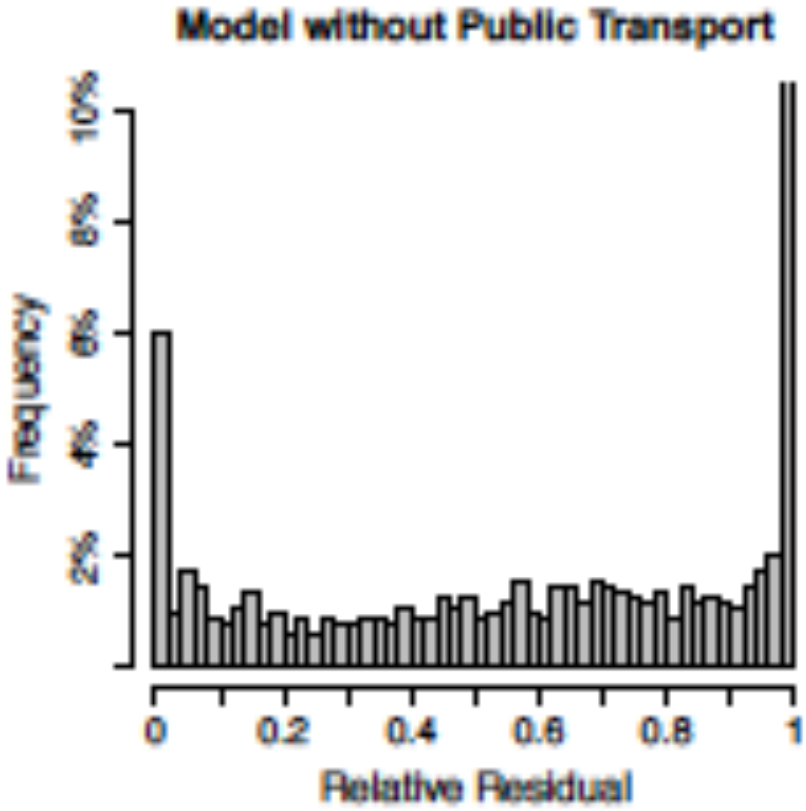
Residuals: False positives of a membership model



Source: Kopp (Forthcomng)

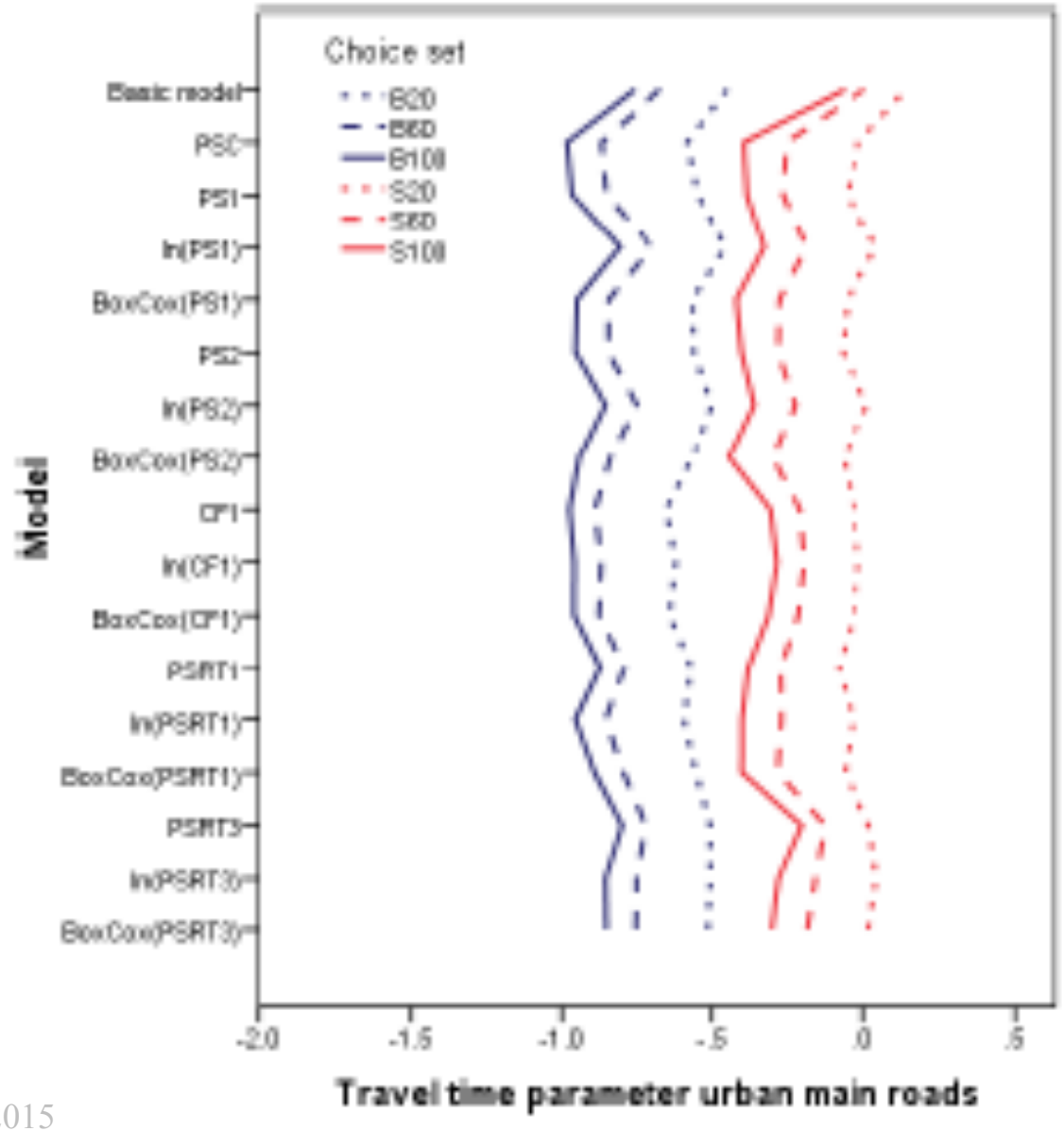
Residuals: MCDEV model of fleet choice





Source: Jäggi (Forthcomng)

Number of non-chosen alternatives: routes



Source: Schüssler (2010)

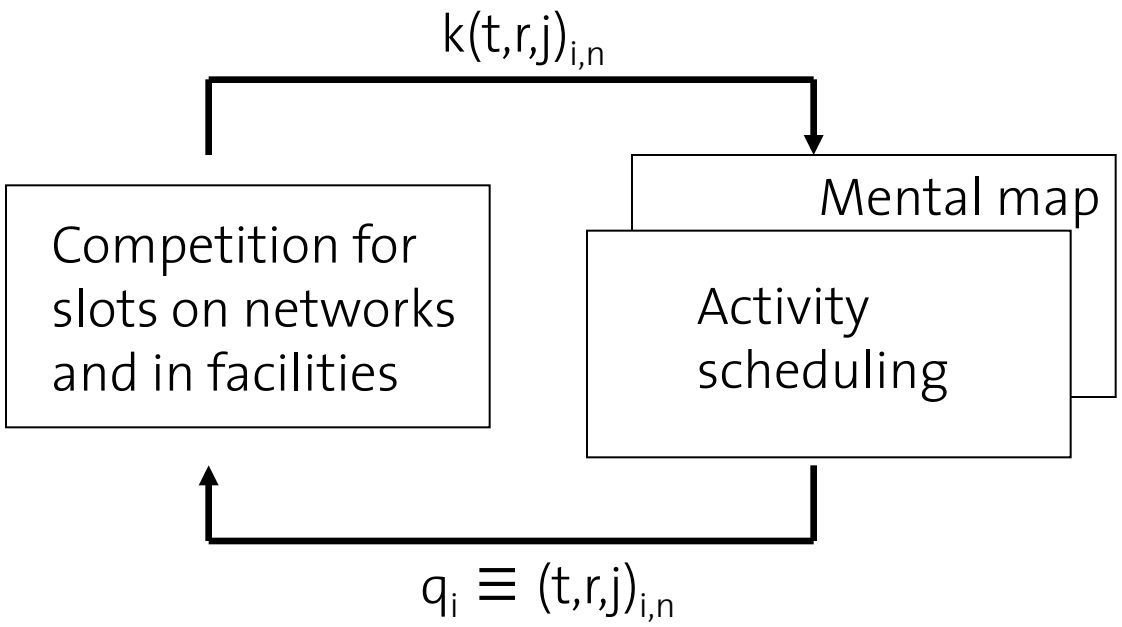
Number of choice sets: residential choice

MEASUREMENTS	ESTIMATES					
	DAT1		DAT2		DAT3	
Household						
DIST_PREVLOC	-5.440	**	-7.070	**	-8.740	**
DIST_WORK	-2.460	*	-3.220	*	-3.880	*
ETA_PREVLOC	0.192	**	0.163	**	0.135	**
ETA_WORK	0.218	**	0.203	**	0.166	**
Accessibility						
MIVACC_CAR	-0.233		-0.302	**	-0.187	
PTACC_NOCAR	0.555	**	0.541	**	0.547	**
Socioeconomic Environment						
SAME_HH_AGE_SHARE	0.782	**	0.684	**	0.634	*
R^2	0.508		0.529		0.524	
adj R^2	0.500		0.522		0.517	

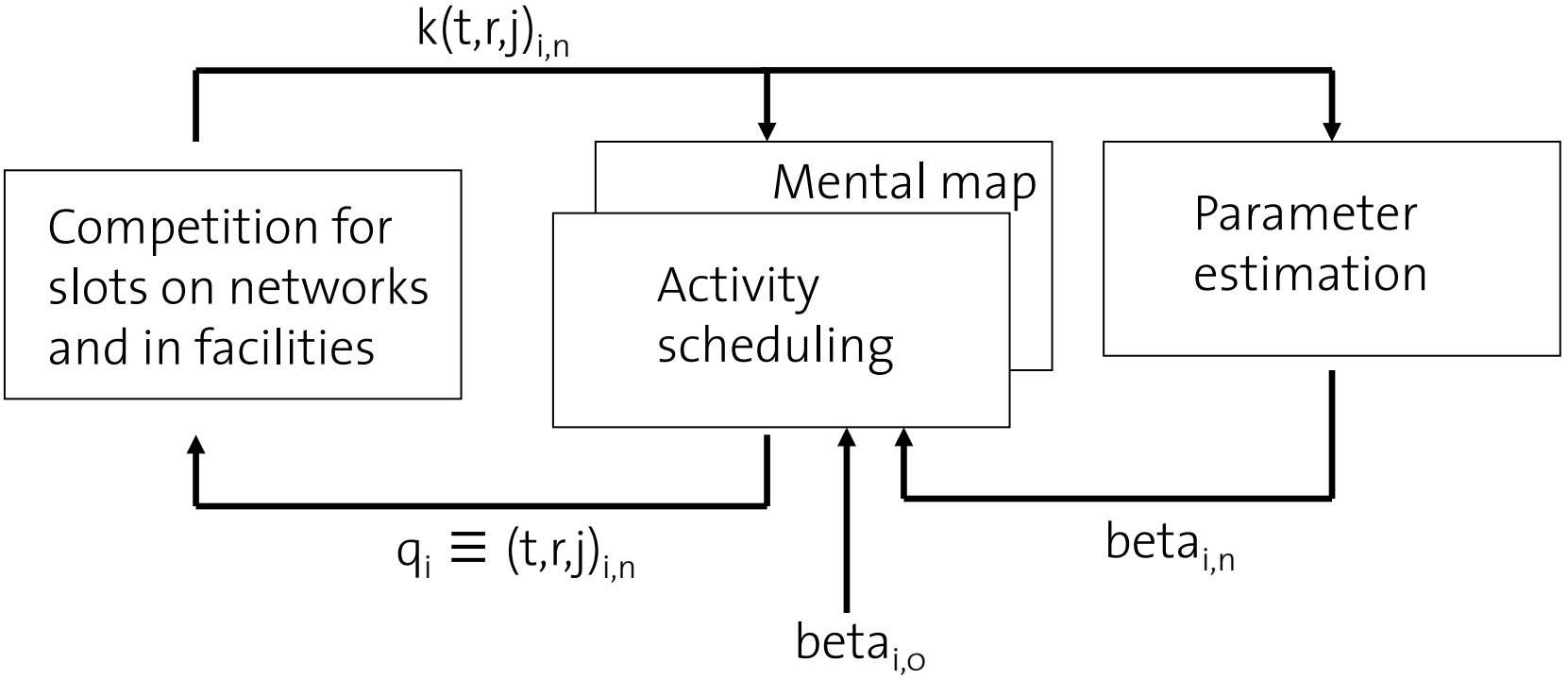
Adapted from: Schirmer (Forthcoming)

Accounting for consistency

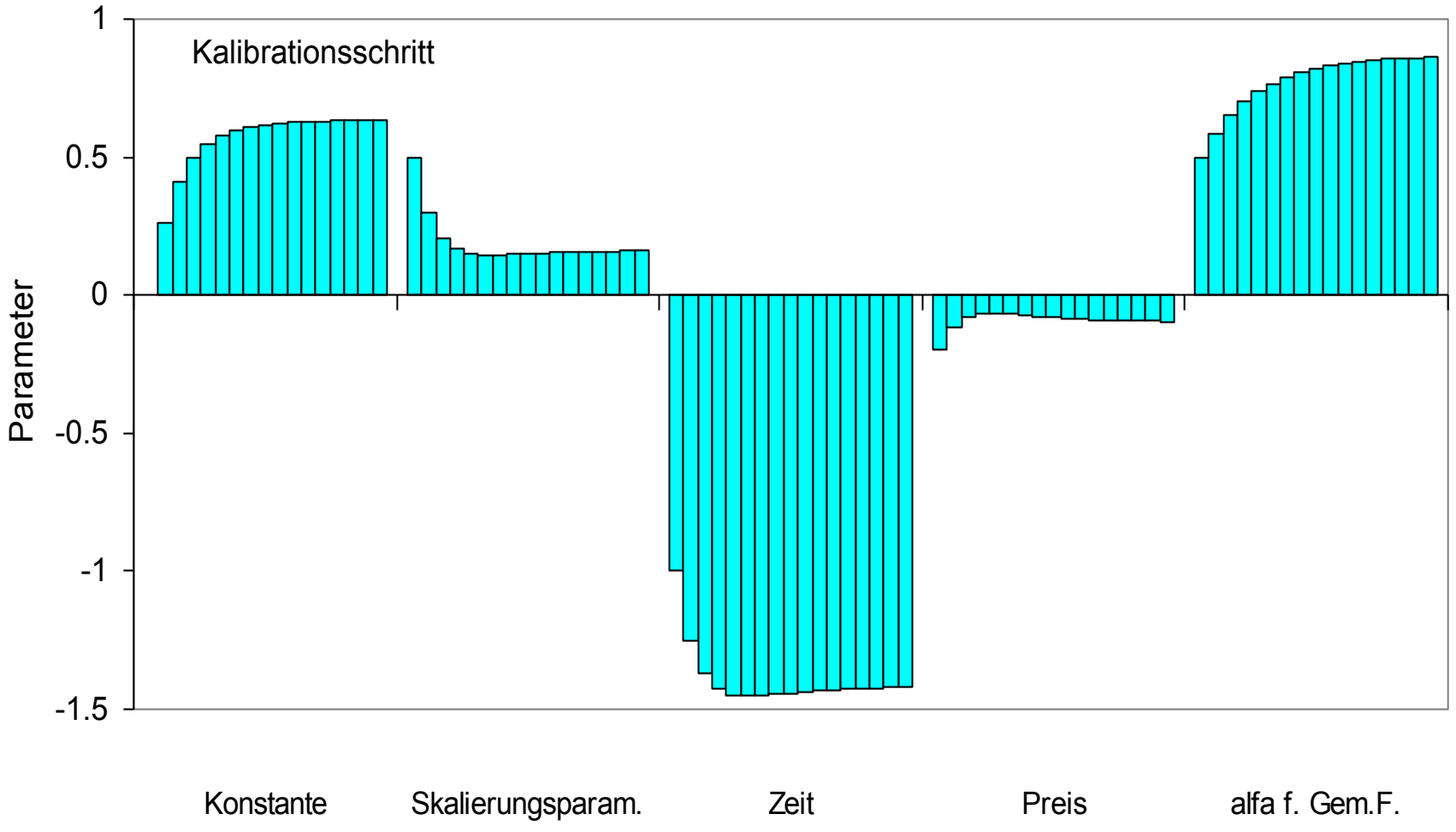
Learning approach of the generic one-day transport model



Model estimation: $beta_{i,o} = beta_{i,n}$? $beta_{i,n-1} = beta_{i,n}$?



Model estimation: $beta_{i,o} = beta_{i,n}$? Route and mode



Source: Vfrtic (2003)

Do we have a MAUP problem ?

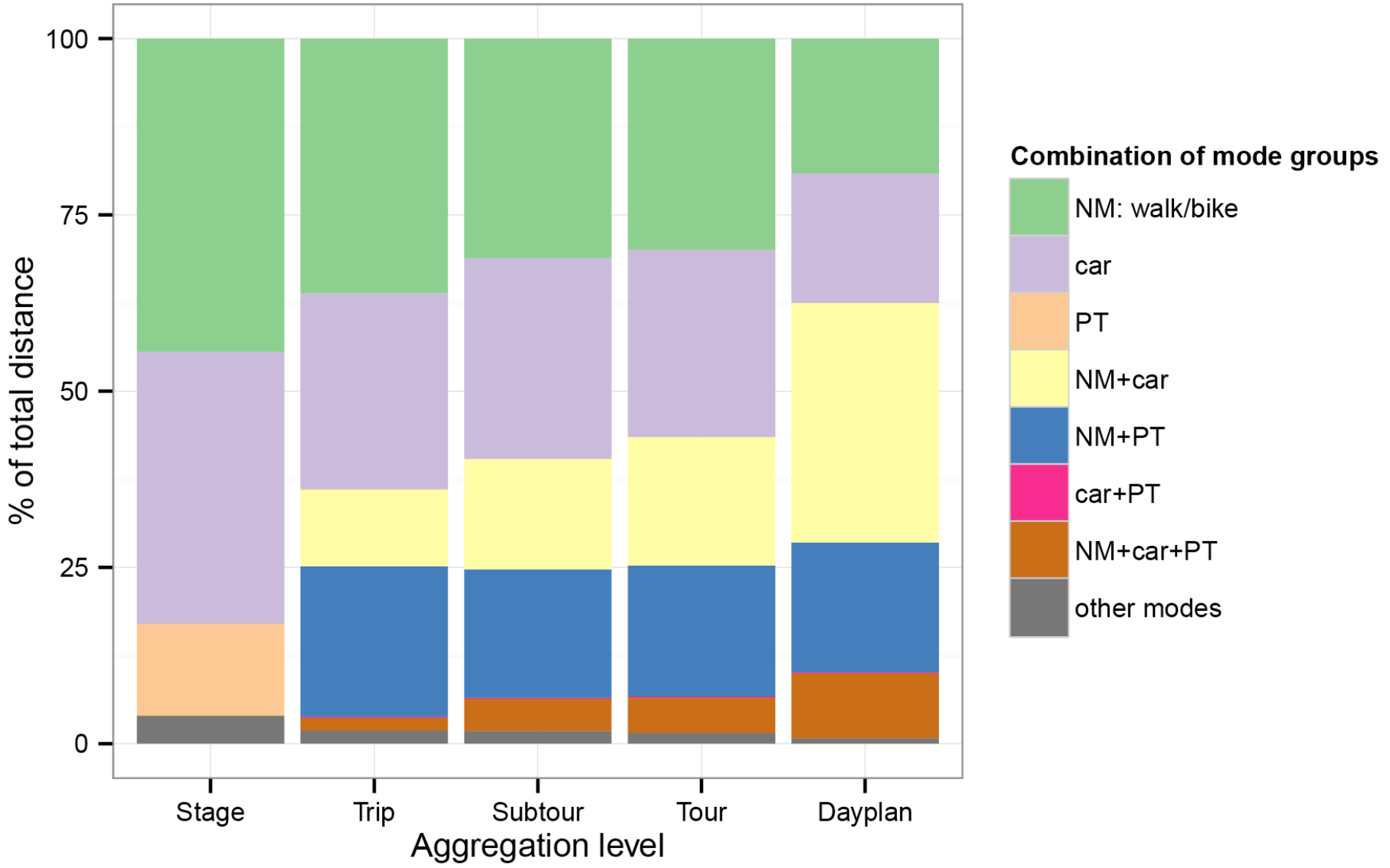
Do we have a MAUP-like problem for DCM?

- Location choice, obviously
- Route choice, obviously
- Time-of-day choice, obviously

- But also, mode choice

    - Stage
    - Trip
    - Sub-tour
    - Tour
    - Daily schedule

Swiss national travel diary 2010: Main mode by aggregation

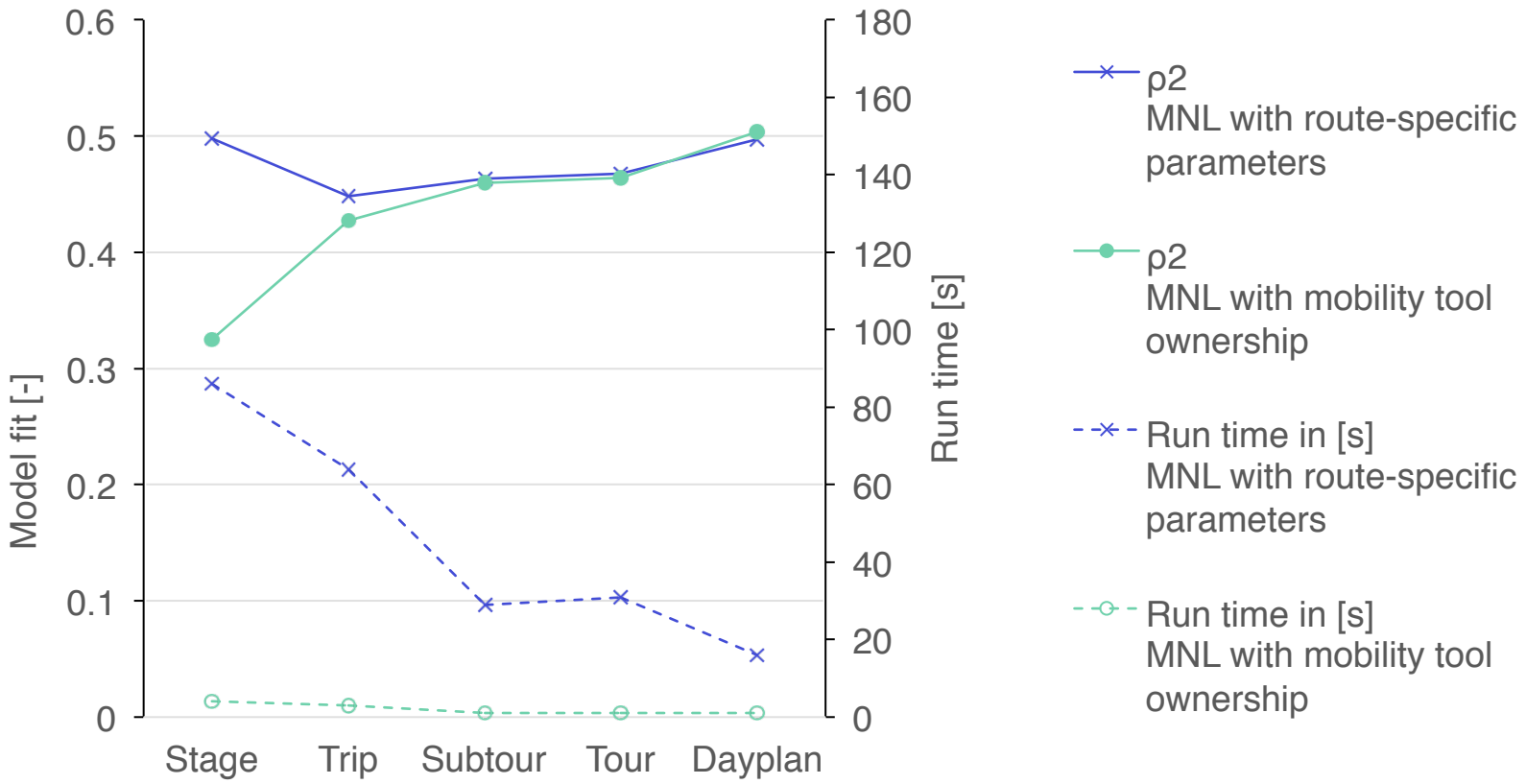


Do we have a MAUP-like problem for DCM?

		Stage	Trip	Subtour	Tour
Value of Time Walking	CHF/h	152	28	26	24
Value of Time Bike	CHF/h	194	39	43	40
Value of Time Car	CHF/h	135	25	30	27
Value of Time PT	CHF/h	-30	2	7	6
Value of Time PT access	CHF/h	819	15	22	22
TT PT / TT Car	-	-4.46	12.33	4.07	4.16
TT Walk / Access time PT	-	0.19	1.83	1.19	1.09
Transfer / TT PT	min	-220.43	107.00	31.28	32.92
Interval / TT PT	-	0.96	7.00	3.47	6.33
Access time / TT PT	-	-27.10	7.67	3.02	3.35

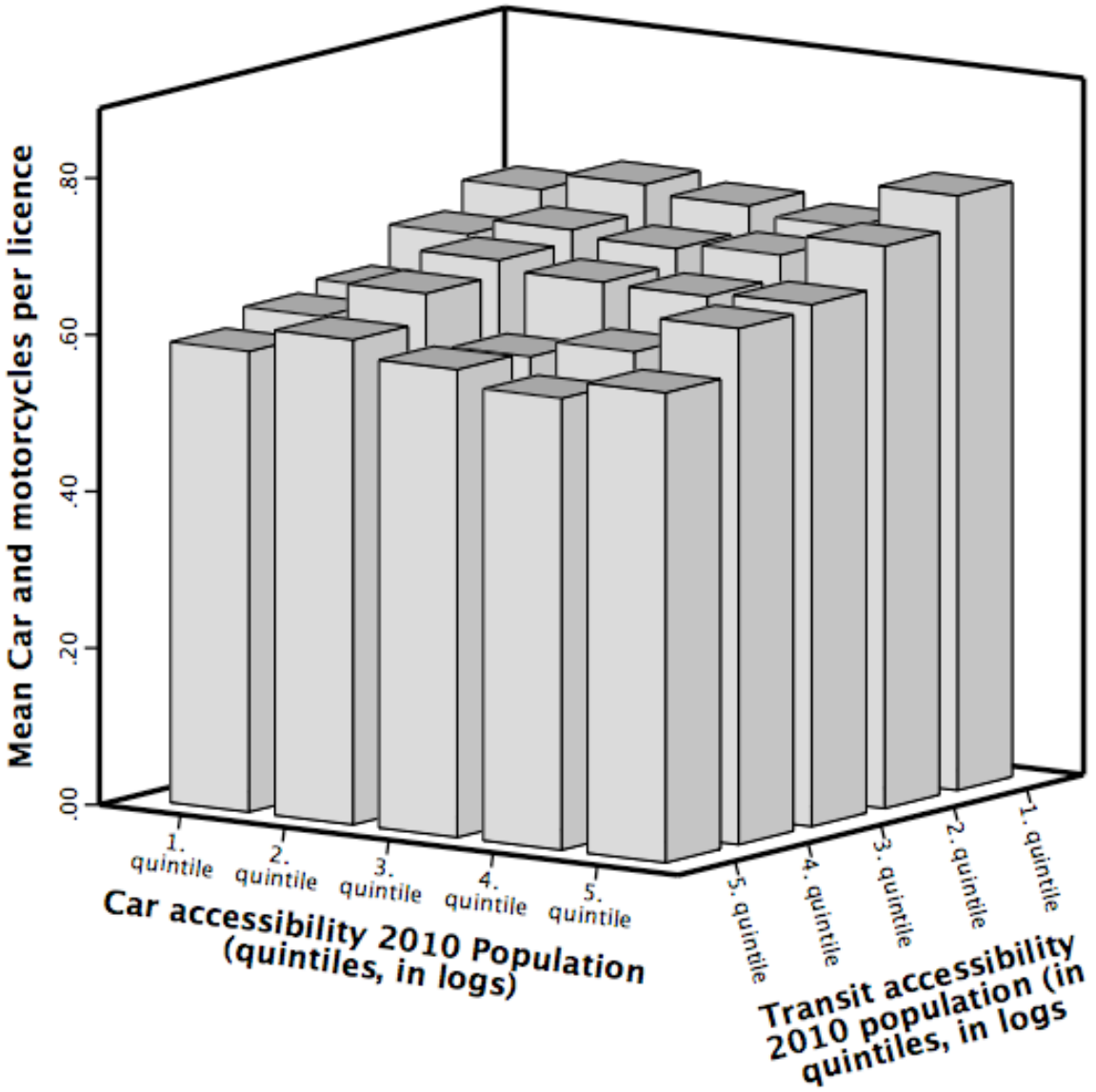
Source: Schmujtz (2015)

Do we have a MAUP-like problem for DCM?



Source: Schmujtz (2015)

Do we get the time horizon right?



Data: Mikrozensus 2010 – Swiss National Travel Diary

What should we do ?

Next steps

- Become more systematic
  - Test for choice set size effects
  - Test for the stability of the estimates wrt choice set
  - Test for the stability wrt imputation of the attribute values
- Check for the right unit of analysis
- Check for the right set of explanatory variables

Questions ?

www.ivt.ethz.ch

References

Jäggi, B. (Forthcoming) Decision modeling on the household level for energy, fleet choice and expenditure, , Dissertation, ETH Zürich, Zürich.

Kopp, J. (2015) GPS-gestützte Evaluation des Mobilitätsverhaltens von free-floating CarSharing-Nutzern, Dissertation, ETH Zürich, Zürich.

Schmutz, Simon (2015) Auswirkung von analytische Einheiten und Aggregationsregeln auf die Verkehrsmittelwahlmodellierung, MSc thesis, Zürich, January 2015.

Schuessler, N. (2010) Accounting for similarities between alternatives in discrete choice models based on high-resolution observations of transport behaviour, ETH Zürich, Zürich.

Vrtic, M. (2003) Simultanes Routen- und Verkehrsmittelwahlmodell, PhD Dissertation, Fakultät für Verkehrswissenschaften, TU Dresden, Dresden.

Zilliak, S. and D. McCloskey (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, University of Michigan Press, Ann Arbor.