Séminaire AME – SPI & SHS, quelle incertitude dans nos modèles?

17 octobre 2018

Calibration and validation of Integrated Transportation and Land Use Models : a survey

NICOLAS COULOMBEL (UNIVERSITÉ PARIS EST, LVMT)

PETER STURM (INRIA, STEEP)

Context

Renewed interest in ITLUM for several years

- imperatives of sustainable development ⇒ need for comprehensive analyses of land use and transport policies
- improvements in computer performance, numerical tools, and data collection address several of Lee's criticisms (Lee, 1973)

The ITLUM literature teems with reviews

- David Simmonds Consultancy et al. (1999), Wegener (2004), …
- But mostly a description (+ analytical comparison) of the models

ITLUM are complex models

- in the processes they are trying to represent
- in their structure

Calibration and validation is a major challenge ⇒ where are we today?

Outline of the presentation

1. Introduction
2. Terminology & Methodology
3. State of the art
4. Conclusion

Calibration: definition

No clear consensus over the exact definition of the term

- view 1: calibration = estimation
- view 2: determine parameters so as to best fit observed data
- view 3: change parameter values (after estimation) based on additional data
- view 4: view 2+ back-and-forths with model design

Our acceptation of the terms:

- calibration = process that determines parameter values to best fit observed data (view 2)
- estimation = use of standard statistical/econometric procedures to determine parameter values

Calibration process: 3 main elements

Calibration strategy (Abraham and Hunt, 2000)

- Limited view
- Piecewise
- Simultaneous
- Sequential
  - one specific instance: Bayesian Sequential

Problem formulation

- Objective function + constraints (prior knowledge)

Solving methods

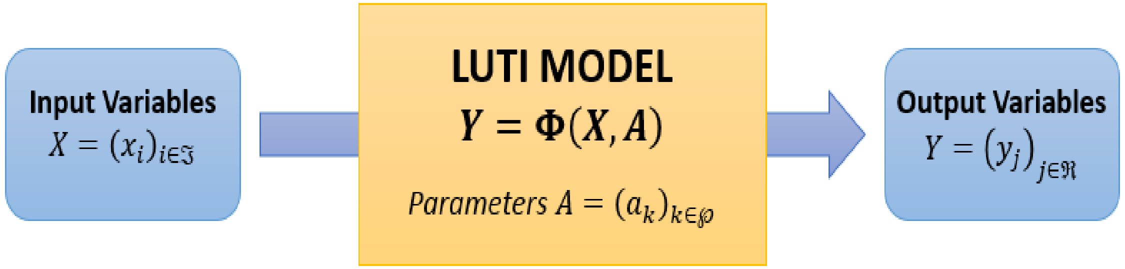
- Numerical tools+ implementation strategies

Limited view strategy

Treat the ITLUM as a black-box and calibrate it all at once

+ The whole calibration procedure is sound in that it aims to reproduce the observations that correspond to the outputs of the modeling exercise
+ Consistency between the calibration and application stages in the way the model is used
+ Possibility to use the reduced form of the model
+ Most likely to reveal structural model deficiencies
– Difficulties relative to the choice of the objective function
– Derivation of the likelihood function will seldom be feasible
– Inability to use additional and/or disaggregate data during calibration

THE ITLUM MODEL IN THE LIMITED VIEW PARADIGM



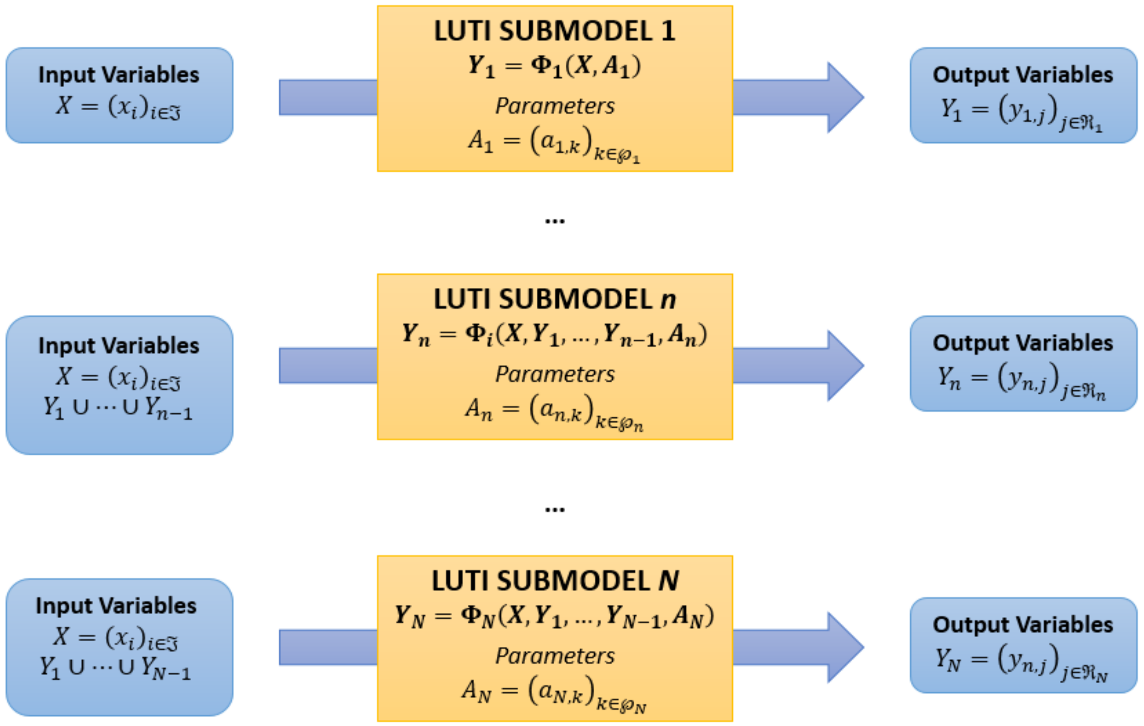
Piecewise strategy

Submodels are calibrated successively and each independently from the others

+ Improved calibration at the submodel level by enabling the use of dedicated estimation methods and extra data
+ Derivation of the likelihood function will often be possible
+ Confidence intervals for the parameters and goodness-of-fit measures will often be available

– Uncertainty regarding the calibration of the ITLUM as a whole
– Inconsistency between the calibration and application stages in the way the model is used
– Absence of comprehensive calibration of the modeling system may lead to several biases, due to systematic errors and/or aggregation biases
– Poor treatment of parameters shared by several submodels

THE ITLUM MODEL IN THE PIECEWISE PARADIGM



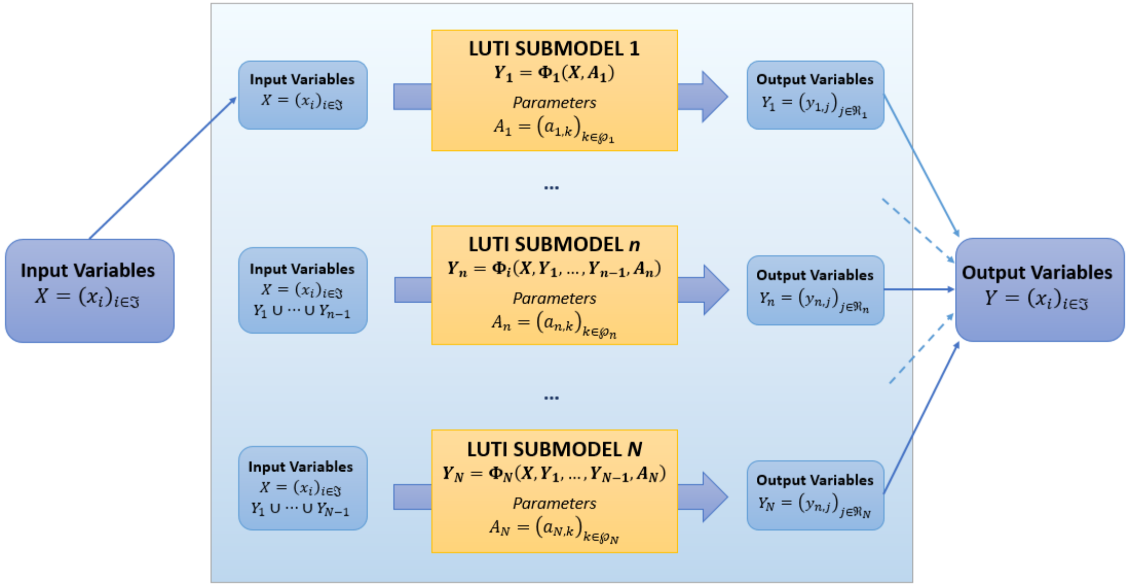
Simultaneous strategy

Combination of the two previous approaches:

simultaneous calibration of each submodel and of the ITLUM as a whole

- + Theoretically pure
- + Combines most of the advantages of the limited view and piecewise strategies
- + Addresses most of the issues of the piecewise strategy
- – Very complex to carry out
- – Difficulties relative to the choice of the objective function for the ITLUM as a whole
- – Difficulties relative to the choice of the composite objective function

THE ITLUM MODEL IN THE SIMULTANEOUS PARADIGM



Sequential strategy

Calibration of each submodel individually, then of the ITLUM as a whole

- Bayesian sequential strategy: statistical information on model parameters in the first step is used as a prior in the second step

\+ Retains most of the advantages of the simultaneous strategy

\+ Simpler to implement

– Difficulties relative to the choice of the objective function for the ITLUM as a whole

– For the parameters that are recalibrated, any statistical information is discarded (save for Bayesian sequential)

Validation: definition

In ITLUM literature, validation often refers to testing the model predictive power

- use of additional data → similar to cross-validation in statistics
  - historical data / additional data sources from the same reference year / split spatial data into two sets: training set vs. testing set

Behavioral validation: from « realism in performance » to « realism in process »

- test of standard policies: urban toll, urban growth boundary, …
- isolating the effect of one or several variables → sensitivity analysis

Uncertainty analysis

- study the propagation of errors in order to quantify uncertainties regarding model outputs

Our acceptation of the term

- Validation = test of the model against its intended usage
- encompasses all three above forms

Typology of indicators

Cross-sectional indicators

Overall/Point value

- Total / Mean
- Stoch : Distribution & confidence interval vs. observed value

Agent population distribution

- Mean + SD
- Plot
- Kolmogorov-Smirnov (K-S) test
- Cross-tabulations

Spatial distribution

- Map / Plot
- R^2 of observed vs. predicted
- Stoch: Coverage indicator
- Stoch: Verification Rank Histogram

Trend indicators

Interperiod variation

- Absolute/Relative

Time series

- Plot

Spatial distribution

- Map / Plot of interperiod variation
- Distribution of difference observed vs. predicted

Model performance indicators

OLS

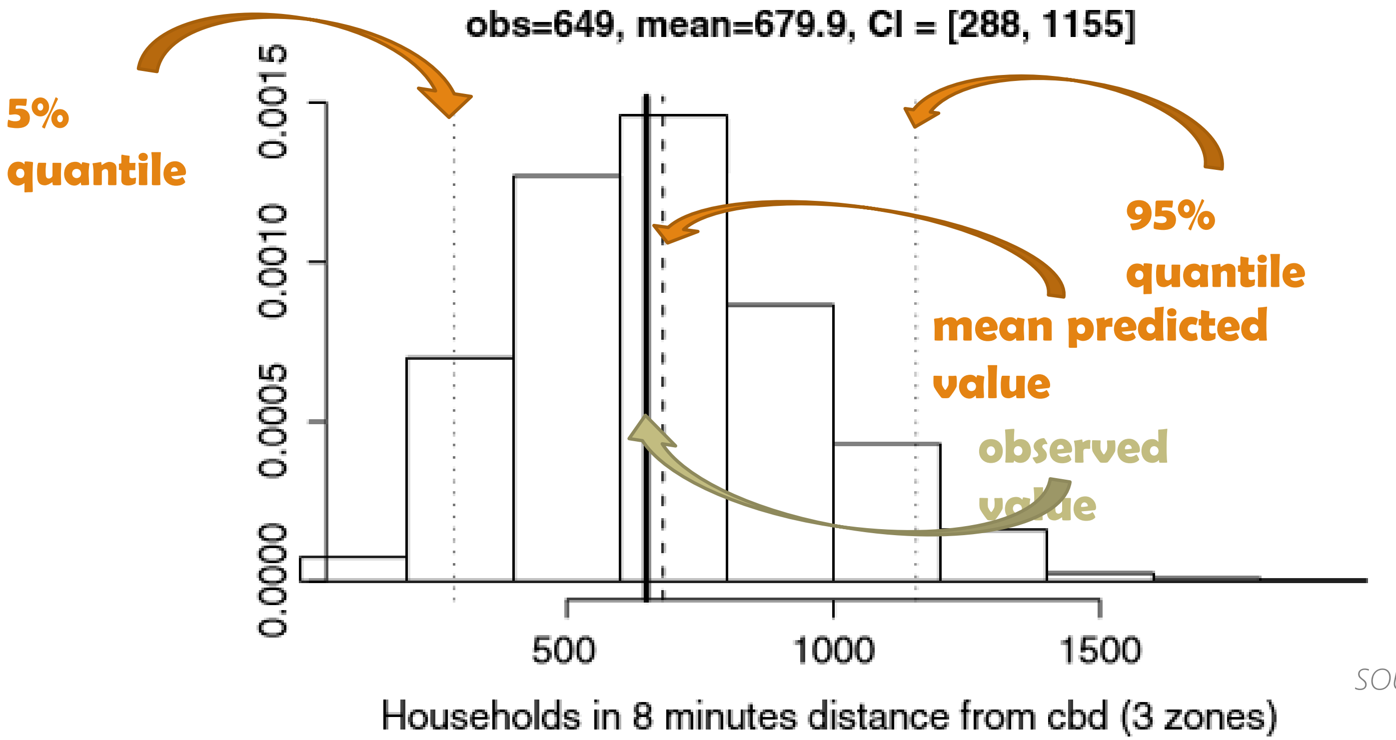
- R^2, Adjusted R^2

Discrete Choice Models

- Pseudo-R^2
- LL, AIC, BIC

Cross-sectional indicators: Overall/Point values

DISTRIBUTION & CONFIDENCE INTERVAL OF PREDICTED VALUES VS. OBSERVED VALUE

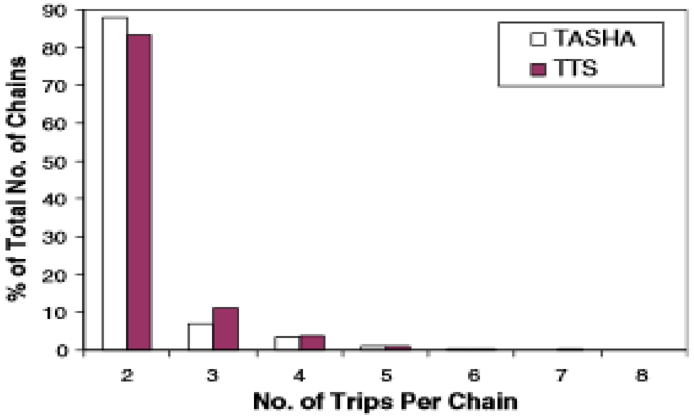


SOURCE: ŠEVČÍKOVÁ ET AL. (2007)

Cross-sectional indicators: Agent population distribution

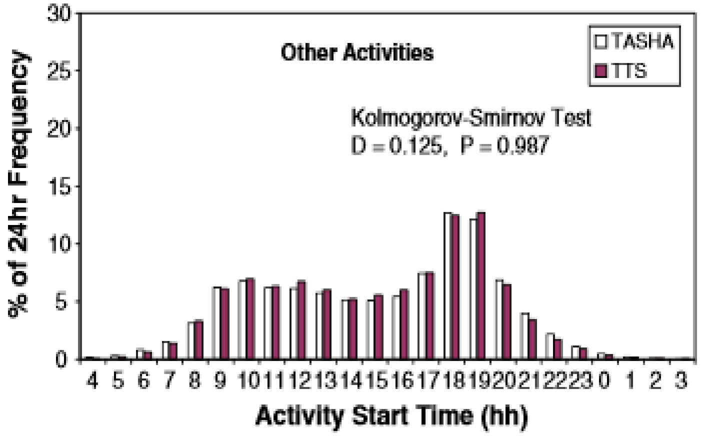
SIMPLE PLOT

SOURCE: ROORDA ET AL., 2008



PLOT + K-S TEST

SOURCE: ROORDA ET AL., 2008



CROSS-TABULATIONS

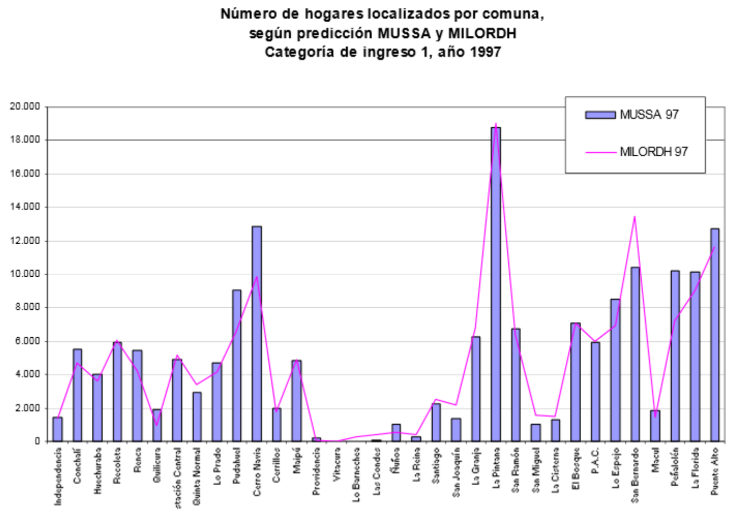
TABLE 1 Observed and Predicted Age Distributions for Married Couples, 2001

Age of Female (years)	Percentage of Couples by Age of Male							
	18–24	25–34	35–44	45–54	55–64	65–74	75–84	85 and older
Census 2001 Married Couples								
18–24	0.28	1.00	0.14	0.03	0.00	0.00	0.00	0.00
25–34	0.18	10.94	7.10	0.39	0.06	0.00	0.00	0.00
35–44	0.02	1.57	19.11	7.84	0.55	0.08	0.00	0.00
45–54	0.01	0.08	1.59	15.21	6.19	0.46	0.03	0.00
55–64	0.00	0.01	0.05	0.95	8.58	4.40	0.24	0.02
65–74	0.00	0.00	0.01	0.04	0.51	5.98	2.39	0.08
75–84	0.00	0.00	0.00	0.00	0.03	0.43	2.56	0.51
85 and older	0.00	0.00	0.00	0.00	0.00	0.01	0.11	0.24
ILUTE 2001 Married Couples								
18–24	1.21	0.71	0.17	0.00	0.00	0.01	0.00	0.00
25–34	0.05	11.40	3.78	1.00	0.03	0.03	0.00	0.00
35–44	0.02	0.97	18.74	8.73	2.38	0.12	0.03	0.00
45–54	0.00	0.40	4.62	12.28	6.32	1.69	0.07	0.00
55–64	0.00	0.01	0.72	3.44	5.96	3.33	0.61	0.02
65–74	0.00	0.01	0.02	0.40	1.83	3.60	1.54	0.23
75–84	0.00	0.00	0.00	0.01	0.10	1.02	1.25	0.47
85 and older	0.00	0.00	0.00	0.00	0.01	0.05	0.28	0.32

SOURCE: MILLER ET AL., 2011

Cross-sectional indicators: Spatial distributions

SIMPLE PLOT



SOURCE: SECTRA – MIDEPLAN (2002)

R^2 OBSERVED VS. PREDICTED

TABLE 1 Goodness-of-Fit of Residential Location Model by Income Group

Parameter	All	Very Poor n=1	Poor n=2	Medium n=3	Medium High n=4	High n=5
R^2	0.75	0.85	0.81	0.77	0.81	0.15

SOURCE: MARTÍNEZ (1996)

COVERAGE INDICATOR

Table 5
Coverage of for the 90% confidence interval

Method	Missed cases	Coverage
Bayesian melding	31	0.88
Multiple runs	163	0.38

Missed cases give the number of observations that fall outside of the confidence interval. The total number of observations is 265.

SOURCE: ŠEVČÍKOVÁ ET AL. (2007)

VERIFICATION RANK HISTOGRAM

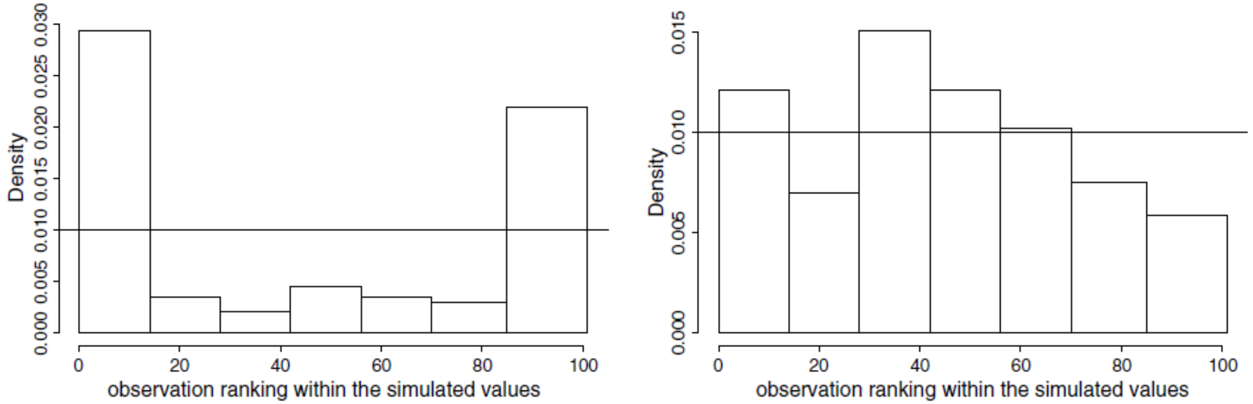


Fig. 7. Verification rank histogram for the output from multiple runs (left panel) and from the Bayesian melding procedure (right panel). The closer the histogram is to being uniform, the better calibrated the corresponding method is.

SOURCE: ŠEVČÍKOVÁ ET AL. (2007)

Trend indicators

INTERPERIOD VARIATION

Table 2
Activity frequency comparison, TASHA vs TTS,

Activity type	Increase in model average distance 1996–2001 (%)	Increase in observed average distance 1996–2001 (%)
Work	6.3	5.8
School	0.0	5.0
Shopping	3.2	11.3
Other	3.1	7.2
Home	5.9	4.8
Total	5.8	5.9

SOURCE: ROORDA ET AL., 2008

TIME SERIES

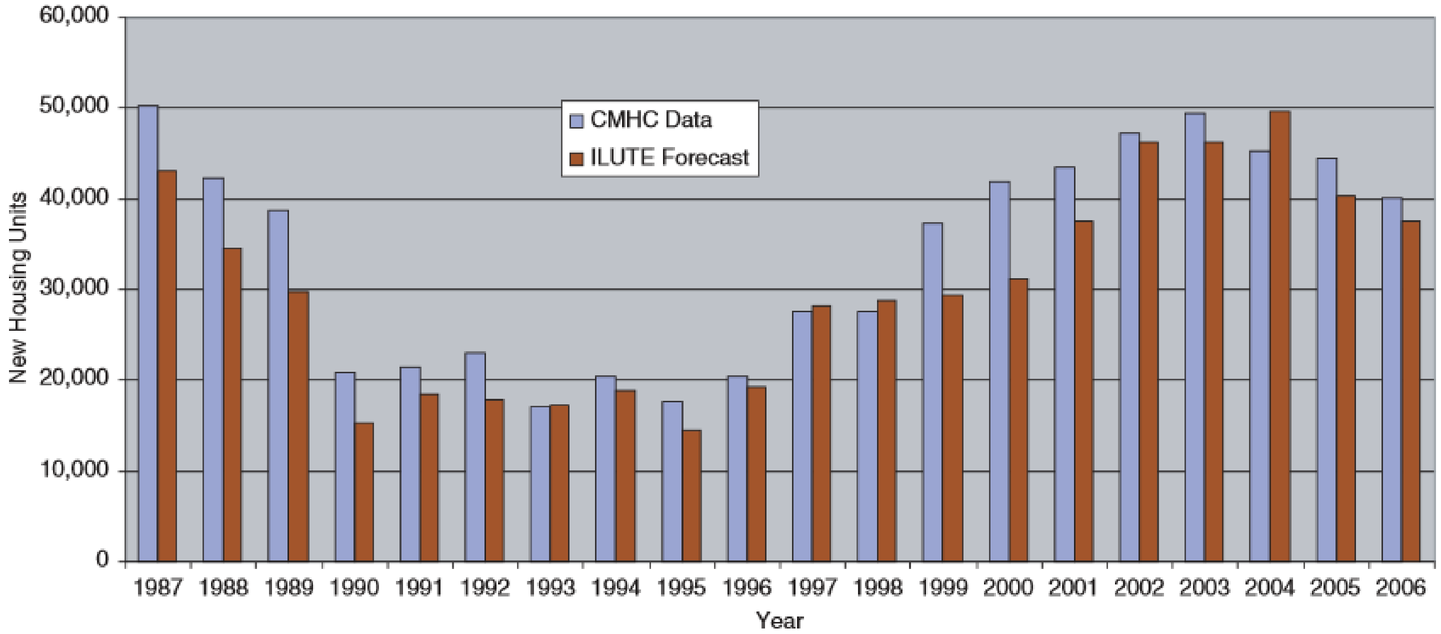


FIGURE 6 Predicted and observed greater Toronto–Hamilton area supply of new housing (CHMC = Canada Mortgage and Housing Corporation). (Source: CHMC.)

SOURCE: MILLER ET AL., 2011

Case of stochastic LUTI models

Outputs are stochastic $\Rightarrow$ point values may not be very informative

Usual strategy:

- run the model N times
- analyze the output distribution
  - often mean – standard deviation (of mean) $\rightarrow \mu_{Runs}$ & σ_{Runs}
  - test whether intrinsic variability of the model results $\lesseqgtr$ difference observed vs. predicted
  - more sophisticated methods: coverage indicator, verification rank histogram, …

Table 2
Activity frequency comparison, TASHA vs TTS, 2001[a]

Activity type	Model average total activities (TASHA)[b]	Model std. dev. total activities (TASHA)[b]	Observed total activities (TTS)
Work	143,990	329	145,123
School	41,987	62	43,930
Shopping	46,844	357	53,989
Other	84,577	360	93,771
Home	26,5031	364	264,588
Total	582,429	1131	601,401

Methodological issue

- Consider not a point value but the distribution of a variable X (age, trip length, house prices, …),
- How do you compute the moments or the distribution of X over N runs?

The *N* runs problem

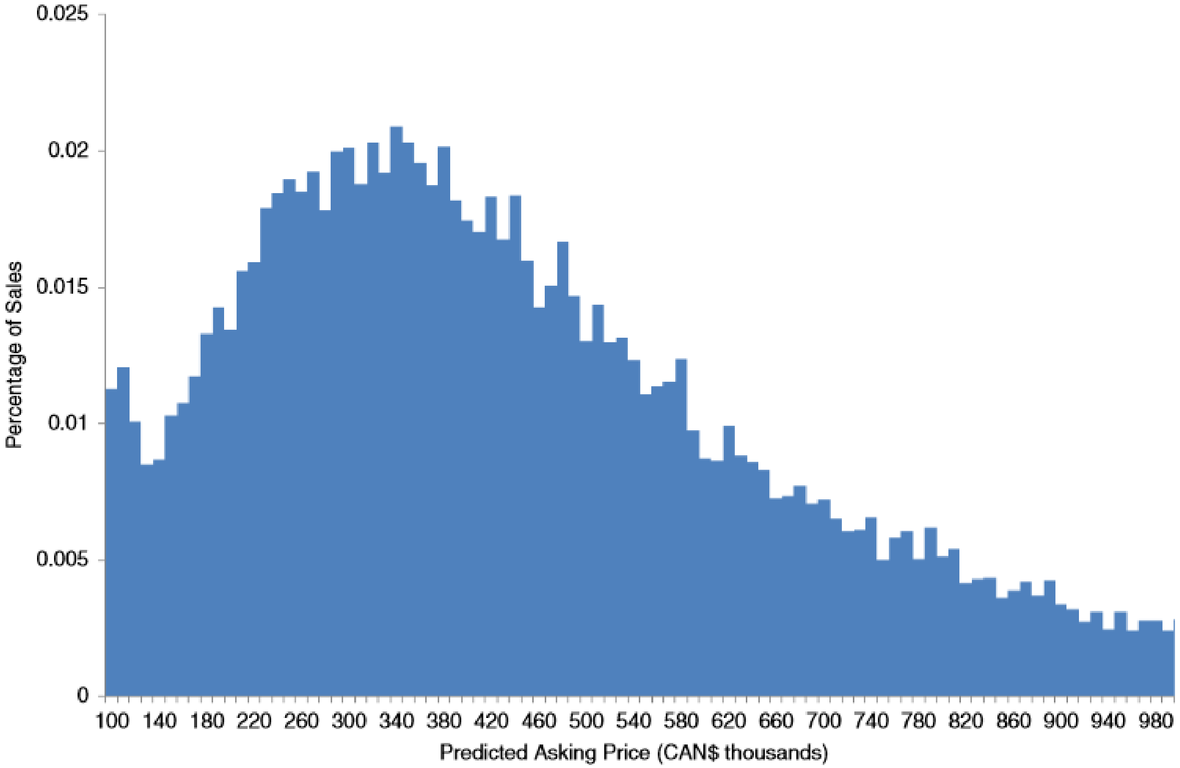


FIGURE 7 Predicted asking prices for housing, 2001.

TABLE 2 Predicted and Observed Transaction Prices by Dwelling Structure Type, 2001

	ILUTE			
Dwelling Type	Average	SD	TREB Average	Delta
Detached	480,000	200,000	307,000	173,000
Semidetached	280,000	130,000	230,000	50,000
Attached	260,000	110,000	212,000	48,000
Apartment	226,000	96,400	182,000	44,000
Total	392,000	180,000	222,000	170,000

NOTE: SD = standard deviation; TREB = Toronto Real Estate Board.

MUSSA – CUBE LAND

Model Type: spatial-economics model

Agent representation: aggregate

Integration Level: standard

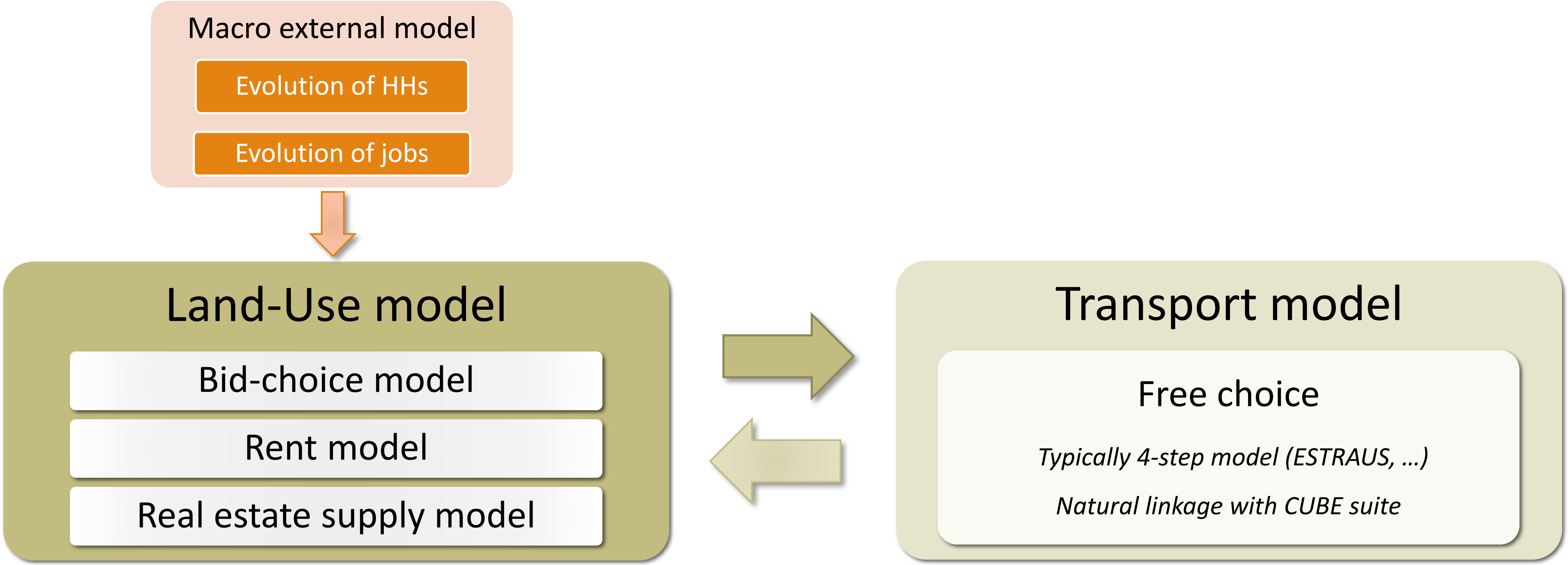
Level of stochasticity

- LU : deterministic
- T : variable *(typically, deterministic)*

Study areas: Santiago (Chile), Montgomery (AL, USA)

Main sources: Martinez (1996), Sectra - Mideplan (2002), Martinez and Donoso (2010), Martinez (2011, PPT)

Typical model structure



Calibration

Strategies

LU – T : separate

LU : piecewise

T : variable

(typically, piecewise)

Methods

LU : standard estimation procedures

- Max LL : bid choice model (MNL), supply model (MNL in aggregate form)
- OLS: rent model

T : variable

Performance Indicators

LU

- Model performance indicators
  - R^2 : rent model
  - Pseudo-R^2 : bid-choice, supply
- Cross-sectional indicators
  - Spatial distribution: R^2 (predicted vs. observed)
    - location of HHs and Firms (**by segment**)

T

- No info

Validation

Historical validation

- Period of analysis: 1991 (calibration year) – 1997 (test year)
- Indicators
  - Cross-sectional indicators
    - Spatial distribution / plot: # of HHs, rents
      - **results per HH segment** (income level)
  - Trend indicators
    - Spatial distribution / plot of inter-period variation : newly-built floor space for economic activities (absolute variation)
- Satisfactory results except for real estate supply model (according to authors)

ILUTE

Model Type: activity-based model

Agent representation: fully disaggregate (with or without sampling)

Integration Level: medium

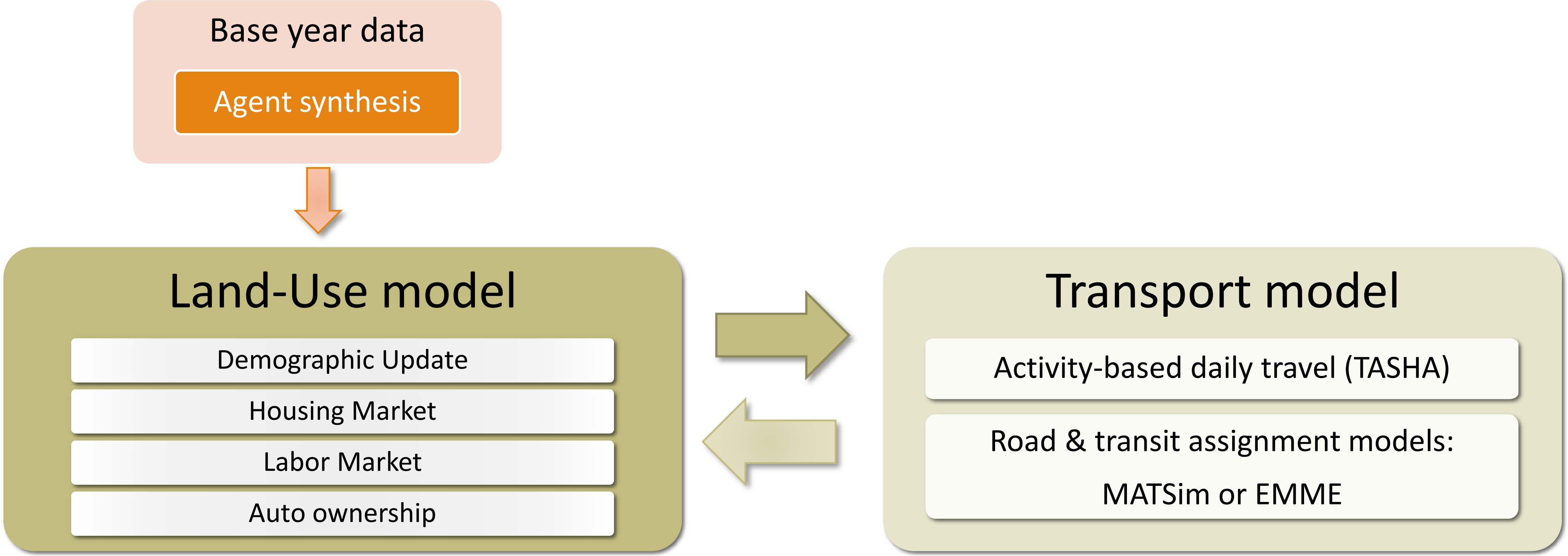
Level of stochasticity: high

- LU : sequence of stochastic submodels
- T : activity scheduling microsimulation model = stochastic & assignment model = variable

Study areas: Greater Toronto-Hamilton area (Canada)

Main sources: Roorda et al. (2008), Miller et al. (2011), Farooq and Miller (2012)

Typical model structure



Calibration

LU – T : separate

LU : piecewise

T : piecewise

Strategies

LU : TO BE COMPLETED

T

- TASHA: standard to advanced estimation procedures
- Assignment models: variable

Methods

LU

- TO BE COMPLETED

T

<u>TASHA</u>

- Stochasticity $\Rightarrow$10 runs $\Rightarrow$ μ_{Runs} & σ_{Runs}
- Cross-sectional indicators
  - Overall/Point Value / μ_{Runs} & σ_{Runs} : # of activities & mean trip length (per activity type)
  - Agent population distribution
    - plot: n° of trips per chain
    - plot + KS test : activity start time & duration

<u>Assignment model</u>

- No info

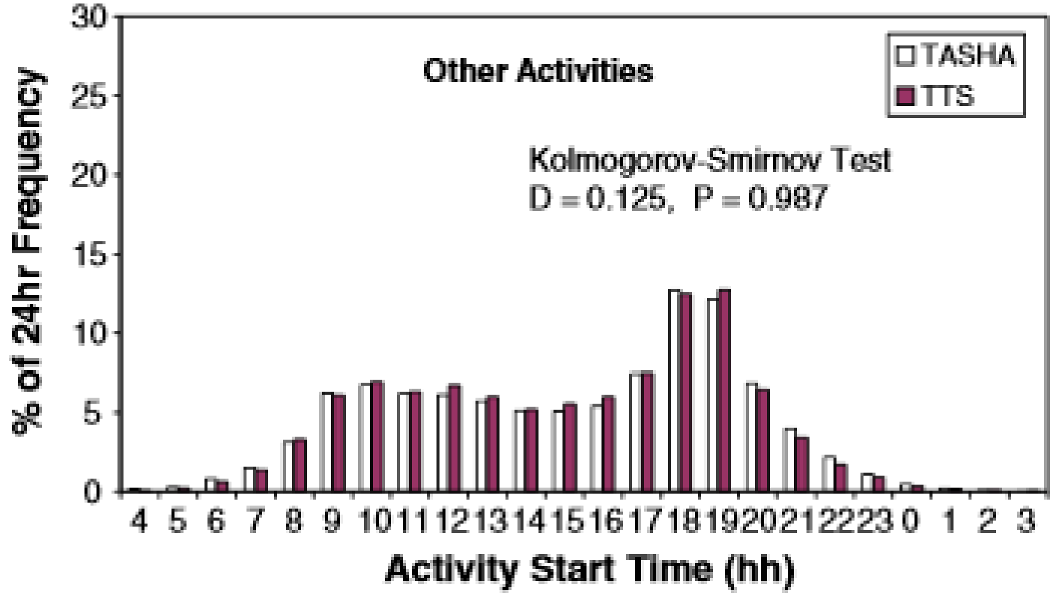
Performance Indicators

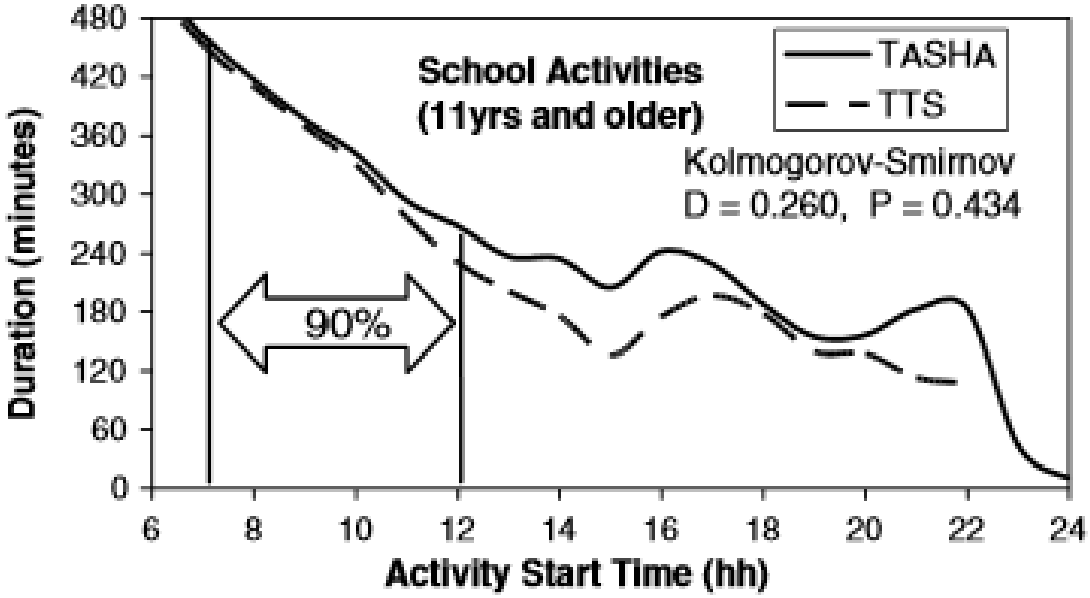
Validation

Historical validation

- 2 validation exercises: 1) TASHA and 2) Part of the land-use submodels
- Period of analysis: 1) 1996 – 2001 and 2) 1986 – 2006
- **Stochasticity taken into account** : 10 runs of ILUTE $\Rightarrow \mu_{Runs}$ & σ_{Runs} (not for all variables)
- Indicators
  - Cross-sectional indicators
    - Overall/Point value :
      - μ_{Runs} : mean trip length (by time of day)
      - μ_{Runs} & σ_{Runs} : # of activities & mean trip length (both by activity type)
    - Agent population distribution :
      - Mean + SD: transaction price (by dwelling structure type)
      - Plot: age of population, income difference between male and female within married couples
      - Plot + KS test: activity start time (by activity type), mean duration by activity start time (by activity type)
      - Cross-tabulations: married couples by age male * age female
  - Trend indicators
    - Time series: births – deaths - out-migrations, new housing units

Some correct and incorrect uses of the KS test





UrbanSim

Model Type: activity-based model

Agent representation: fully disaggregate (with or without sampling)

Integration Level: standard to medium (depending on transport model)

Level of stochasticity: high

- LU : sequence of stochastic submodels
- T : variable

Study areas: Eugene-Springfield (OR, USA), Puget Sound Region (WA, USA) Austin (TX, USA), Paris (France), Lyon (France), Brussels (Belgium)…

Main sources: Waddell (2002, 2011), Pradhan and Kockelman (2002), Ševčíková et al. (2007, 2011), Patterson et al. (2010), Kakaraparthi and Kockelman (2011)

Typical model structure

Base year data

Agent synthesis

Macro external model

Evolution of HHs

Evolution of jobs

Land-Use model

- Economic and Demographic Transition Models
- Employment and Household Mobility Model
- Employment and Household Location Model
- Land Price Model
- Real Estate Development Model

Transport model

Free choice

Activity-based transport model with DTA (MATSim, …)

Standard static 4-step model (EMME, TransCAD, …)

Calibration

Strategies

LU – T : separate

LU : piecewise in most applications

- **Bayesian sequential** : based on Bayesian melding (Ševčíková et al., 2007, 2011)

T : variable

(typically, piecewise)

Methods

LU : standard estimation procedures (mostly)

- Mobility models: random sampling → observed mobility rates (sample mean)
- Location choice models: MNL → max LL
- Land price model: hedonic model → OLS
- Real estate development model: MNL → max LL

T : variable

Performance Indicators

LU

- Model performance indicators
  - R^2 : land price model
  - Pseudo-R^2 : location choice models, real estate development model
- Cross-sectional indicators (Bayesian melding)
  - Overall/Point value
    - distribution & confidence interval vs. observed value: # of HHs in one specific zone
  - Spatial distribution
    - coverage indicator: # of HHs per zone
    - verification rank histogram: # of HHs per zone

T : variable

Validation (1)

Historical validation

- Waddell (2002): a "pseudo-instance" of historical validation
- Period of analysis: 1980 (start year) – 1994 (calibration & test year)
- Indicators
  - Cross-sectional indicators
    - Spatial distribution / correlation of observed to predicted (⇔ R^2) : employment, population, non residential sq feet, housing units, land price
      - **results for 3 spatial levels** (cell, average over 1 cell radius, zone)
  - Trend indicators
    - Spatial distribution / distribution of difference observed vs. predicted: employment & population (per zone)

Validation (2)

Sensitivity analysis

- scenarios: bridge construction (Nicolai et al., 2011), range of 6 transport and/or land use scenarios (Kakaraparthi and Kockelman, 2011)
  - Indicators (Nicolai et al., 2011):
    - Trend indicators / Time series / Plot : travel time to Seattle CBD, accessibility to jobs,(# of jobs within 30 minutes), housing prices, population (in Bainbridge), # of single-family (including vacant) and multi-family residential units
  - Indicators (Kakaraparthi and Kockelman, 2011)
    - Cross-sectional indicators / Overall/Point value: daily VMT, average speed, mean V/C ratio, average HH and job density, average HH and job accessibility, energy consumption (per sector)
    - Cross-sectional indicators / Spatial distribution / Maps : HH and job densities
  - no clear expectations in Nicolai et al. (2011) vs. ad verecundiam (argument from authority) in Kakaraparthi and Kockelman (2011)
  - stochasticity of the ITLUM not accounted for

Validation (3)

Uncertainty propagation

- Factorized design approach (Pradhan and Kockelman, 2002)
  - Considers uncertainties in model input and model parameters → 81 scenarios
  - Analysis of impact of uncertainties by regressing output on inputs / parameters and use of standardized coefficients
  - Output variables: LU (housing prices, occupancy rate & density) & T (VMT, VHT, flows on 3 main road links)
  - Short comparison with intrinsic stochasticity of the model (appraised with 10 runs)
- Bayesian melding (Ševčíková et al., 2007, 2011)
  - Theoretical framework developed to consider both uncertainties linked to model inputs / parameters & to stochasticity of the ITLUM
  - Random draws of model parameters and input variables
    - Model parameters: distribution based on estimation results at t_0
    - Input variables: distribution based on variability of several independent forecasts
  - Uses intermediate information at t_1 to improve calibration + measure model uncertainty
  - Provides posterior distributions for output variables at t_2 (and for model parameters)
  - Output variable: only LU (# of HHs per TAZ)

Some methodological issues

R^2 predicted vs. observed

- Assume you always predict half the true value $\Rightarrow R^2 = 1$ even though your model is wrong…

Stochastic ITLUM: already discussed

Comparison with a benchmark: naïve model (past trend, …)

What are the relevant indicators?

- LUTI models aim to predict the spatial dynamics of a system
  - Trend indicators should often be preferred to cross-sectional indicators, especially for extensive variables (population, housing)
- Think about the submodels involved
  - Analysis of # of HHs per zone: without segmentation, it mainly tests the supply model, with segmentation, you truly test all models

Preliminary conclusions (1)

No consensus in how to calibrate and validate ITLUMs

- strategies, methods and indicators strongly vary from one case to the other and are seldom justified
- often driven by data availability and model structure

Calibration

- LU and T are always calibrated separately
- Piecewise strategy largely prevails
  - Few instances of sequential strategy: standard or Bayesian sequential
  - For now, no instance of black-box strategy or of simultaneous strategy
- Use of prior knowledge: very rare under the form of parameter constraint, sometimes done by hand (expert say)

Preliminary conclusions (2)

Validation

- Historical validation is the most common form of validation
  - choice of indicators not always relevant (cf. slide « Some Methodological Issues »)
  - no comparison to a benchmark model: could help in the assessment of the quality of the results
- Sensitivity analyses are also relatively frequent
  - mostly under the form of scenarios
  - critical issue: defining scenarios for which the expected effects are well-known and solid
- Uncertainty propagation exercises remain pretty rare
  - Bayesian melding seems especially promising in contributing both to model calibration and validation

Still preliminary conclusions

- very long process as information is spread across papers and technical reports
- objectives: survey of common practices, try and identify good practices and promising methods, aim for some normalization?