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 acceptance 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
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
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

**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

**Interperiod variation**
- Absolute/Relative

**Time series**
- Plot

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

**Trend indicators**

**Model performance indicators**

**OLS**
- $R^2$, Adjusted $R^2$

**Discrete Choice Models**
- Pseudo-$R^2$
- LL, AIC, BIC

---

**Cross-sectional indicators**
Cross-sectional indicators:
Overall/Point values

DISTRIBUTION & CONFIDENCE INTERVAL OF PREDICTED VALUES VS. OBSERVED VALUE

obs=649, mean=679.9, CI = [288, 1155]

Cross-sectional indicators: Agent population distribution

Table 1: Observed and Predicted Age Distributions for Married Couples, 2001

<table>
<thead>
<tr>
<th>Age of Female (years)</th>
<th>18–24</th>
<th>25–34</th>
<th>35–44</th>
<th>45–54</th>
<th>55–64</th>
<th>65–74</th>
<th>75–84</th>
<th>85 and older</th>
</tr>
</thead>
<tbody>
<tr>
<td>Census 2001 Married Couples</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18–24</td>
<td>0.28</td>
<td>1.00</td>
<td>0.14</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>25–34</td>
<td>0.18</td>
<td>10.04</td>
<td>7.10</td>
<td>0.39</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>35–44</td>
<td>0.02</td>
<td>1.57</td>
<td>19.11</td>
<td>7.84</td>
<td>0.55</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>45–54</td>
<td>0.01</td>
<td>0.08</td>
<td>1.59</td>
<td>15.21</td>
<td>6.19</td>
<td>0.46</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>55–64</td>
<td>0.00</td>
<td>0.01</td>
<td>0.05</td>
<td>0.95</td>
<td>8.38</td>
<td>4.40</td>
<td>0.24</td>
<td>0.02</td>
</tr>
<tr>
<td>65–74</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.04</td>
<td>0.51</td>
<td>5.98</td>
<td>2.39</td>
<td>0.08</td>
</tr>
<tr>
<td>75–84</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>0.43</td>
<td>2.56</td>
<td>0.51</td>
</tr>
<tr>
<td>85 and older</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.11</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Kolmogorov-Smirnov Test
D = 0.125, P = 0.067

Source: Miller et al., 2011
Cross-sectional indicators: Spatial distributions

**SIMPLE PLOT**

Número de hogares localizados por comuna, según predicción MOCHA y MIORCH

Categoría de ingreso 1, año 1997

**R² OBSERVED VS. PREDICTED**

**COVERAGE INDICATOR**

Table 5
Coverage of for the 90% confidence interval

<table>
<thead>
<tr>
<th>Method</th>
<th>Missed cases</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian melding</td>
<td>31</td>
<td>0.88</td>
</tr>
<tr>
<td>Multiple runs</td>
<td>163</td>
<td>0.38</td>
</tr>
</tbody>
</table>

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

**VERIFICATION RANK HISTOGRAM**

**TABLE 1**

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>All</th>
<th>Very</th>
<th>Poor n=1</th>
<th>Poor n=2</th>
<th>Medium n=3</th>
<th>Medium High n=4</th>
<th>High n=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>R²</td>
<td>0.75</td>
<td>0.83</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>

**SOURCE:** MARTÍNEZ (1996)

**SOURCE:** SECTRA – MIDEPLAN (2002)

**SOURCE:** ŠEVČÍKOVÁ ET AL. (2007)
Trend indicators

INTERPERIOD VARIATION

Table 2
Activity frequency comparison, TASHA vs TTS.

<table>
<thead>
<tr>
<th>Activity type</th>
<th>Increase in model average distance 1996–2001 (%)</th>
<th>Increase in observed average distance 1996–2001 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work</td>
<td>6.3</td>
<td>5.8</td>
</tr>
<tr>
<td>School</td>
<td>0.0</td>
<td>5.0</td>
</tr>
<tr>
<td>Shopping</td>
<td>3.2</td>
<td>11.3</td>
</tr>
<tr>
<td>Other</td>
<td>3.1</td>
<td>7.2</td>
</tr>
<tr>
<td>Home</td>
<td>5.9</td>
<td>4.8</td>
</tr>
<tr>
<td>Total</td>
<td>5.8</td>
<td>5.9</td>
</tr>
</tbody>
</table>

SOURCE: ROORDA ET AL., 2008

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 ⇒ point values may not be very informative

Usual strategy:

◦ run the model \( N \) times
◦ analyze the output distribution
  ◦ often mean – standard deviation (of mean) → \( \mu_{\text{Runs}} \) & \( \sigma_{\text{Runs}} \)
  ◦ test whether intrinsic variability of the model results ≤ difference observed vs. predicted
  ◦ more sophisticated methods: coverage indicator, verification rank histogram, ...

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?

<table>
<thead>
<tr>
<th>Activity type</th>
<th>Model average total activities (TASHA)(^b)</th>
<th>Model std. dev. total activities (TASHA)(^b)</th>
<th>Observed total activities (TTS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work</td>
<td>143,990</td>
<td>329</td>
<td>145,123</td>
</tr>
<tr>
<td>School</td>
<td>41,987</td>
<td>62</td>
<td>43,930</td>
</tr>
<tr>
<td>Shopping</td>
<td>46,844</td>
<td>357</td>
<td>53,989</td>
</tr>
<tr>
<td>Other</td>
<td>84,577</td>
<td>360</td>
<td>93,771</td>
</tr>
<tr>
<td>Home</td>
<td>26,5031</td>
<td>364</td>
<td>264,588</td>
</tr>
<tr>
<td>Total</td>
<td>582,429</td>
<td>1131</td>
<td>601,401</td>
</tr>
</tbody>
</table>
The $N$ runs problem

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

<table>
<thead>
<tr>
<th>Dwelling Type</th>
<th>ILUTE Average</th>
<th>ILUTE SD</th>
<th>TREB Average</th>
<th>Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detached</td>
<td>480,000</td>
<td>200,000</td>
<td>307,000</td>
<td>173,000</td>
</tr>
<tr>
<td>Semi-detached</td>
<td>280,000</td>
<td>130,000</td>
<td>230,000</td>
<td>50,000</td>
</tr>
<tr>
<td>Attached</td>
<td>260,000</td>
<td>110,000</td>
<td>212,000</td>
<td>48,000</td>
</tr>
<tr>
<td>Apartment</td>
<td>226,000</td>
<td>96,400</td>
<td>182,000</td>
<td>44,000</td>
</tr>
<tr>
<td>Total</td>
<td>392,000</td>
<td>180,000</td>
<td>222,000</td>
<td>170,000</td>
</tr>
</tbody>
</table>

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

**FIGURE 7**  Predicted asking prices for housing, 2001.
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

Macro external model
- Evolution of HHs
- Evolution of jobs

Land-Use model
- Bid-choice model
- Rent model
- Real estate supply model

Transport model
- Free choice
  - Typically 4-step model (ESTRAUS, ...)
  - Natural linkage with CUBE suite
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

Base year data
Agent synthesis

Land-Use model
Demographic Update
Housing Market
Labor Market
Auto ownership

Transport model
Activity-based daily travel (TASHA)
Road & transit assignment models:
MATSim or EMME
Calibration

**Strategies**

- **LU – T**: separate
  - **LU**: piecewise
  - **T**: piecewise

**Methods**

- **LU**: TO BE COMPLETED
  - **TASHA**: standard to advanced estimation procedures
  - **Assignment models**: variable

**Performance Indicators**

- **LU**: TO BE COMPLETED
  - **TASHA**
    - Stochasticity \(\Rightarrow 10 \text{ runs} \Rightarrow \mu_{\text{Runs}} \& \sigma_{\text{Runs}}\)
    - Cross-sectional indicators
      - Overall/Point Value / \(\mu_{\text{Runs}} \& \sigma_{\text{Runs}}\): # of activities & mean trip length (per activity type)
      - Agent population distribution
        - plot: n° of trips per chain
        - plot + KS test: activity start time & duration
      - **Assignment model**
        - No info
Validation

Historical validation

- 2 validation exercises: 1) TASHA and 2) Part of the land-use submodels
- **Stochasticity taken into account**: 10 runs of ILUTE \( \mu_{\text{Runs}} \) & \( \sigma_{\text{Runs}} \) (not for all variables)
- Indicators
  - Cross-sectional indicators
    - Overall/Point value:
      - \( \mu_{\text{Runs}} \): mean trip length (by time of day)
      - \( \mu_{\text{Runs}} \) & \( \sigma_{\text{Runs}} \): # 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)...

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, ...)

Introduction
Terminology & Methodology
State of the art
Conclusion
Calibration

LU – T: separate

LU: piecewise in most applications
• Bayesian sequential: based on Bayesian melding (Ševčíková et al., 2007, 2011)

T: variable
(typically, piecewise)

Strategies

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

Methods

LU
• Model performance indicators
  ○ R²: land price model
  ○ Pseudo-R²: 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

Performance Indicators
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 ($\Leftrightarrow R^2$): 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 $\rightarrow$ 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

R² predicted vs. observed

- Assume you always predict half the true value ⇒ R² = 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?