Generating population data for activity-based travel analysis

Presentation to MOEBIUS Conference,
October 2013

David Simmonds
David Simmonds Consultancy and
Heriot-Watt University

DAVID SIMMONDS CONSULTANCY
URBAN, REGIONAL AND TRANSPORT PLANNING

Why generate synthetic population data?

- Data for the base year(s) has to be synthesized because, even if suitable data for a large or 100% sample exists in Census data, we are not allowed to access it; plus
  - activity-based modelling may well need data which is not available in the Census
- Data for future years has to be synthesized because it cannot yet be observed
    - note that we are dealing with models that typically require 100% samples; even if they can be run with <100% they require a random sample rather than a stratified sample with expansion factors.

Household and person variables

Most of the household variables required are determined by the combination of person characteristics; exceptions are

- tenure (and dwelling characteristics)
- car ownership
- household income (where tax/benefit rules mean that household income is not simply the sum of independently determined individual incomes).

Synthesizing base year data

A range of methods exists which

- take individual cases from a suitable sample survey of households and household members (or sample data from a 100% survey), with limited geographical data
- duplicates them and allocates them to more specific or very specific locations (in some cases, to an individual dwelling).

Two points to note about this:

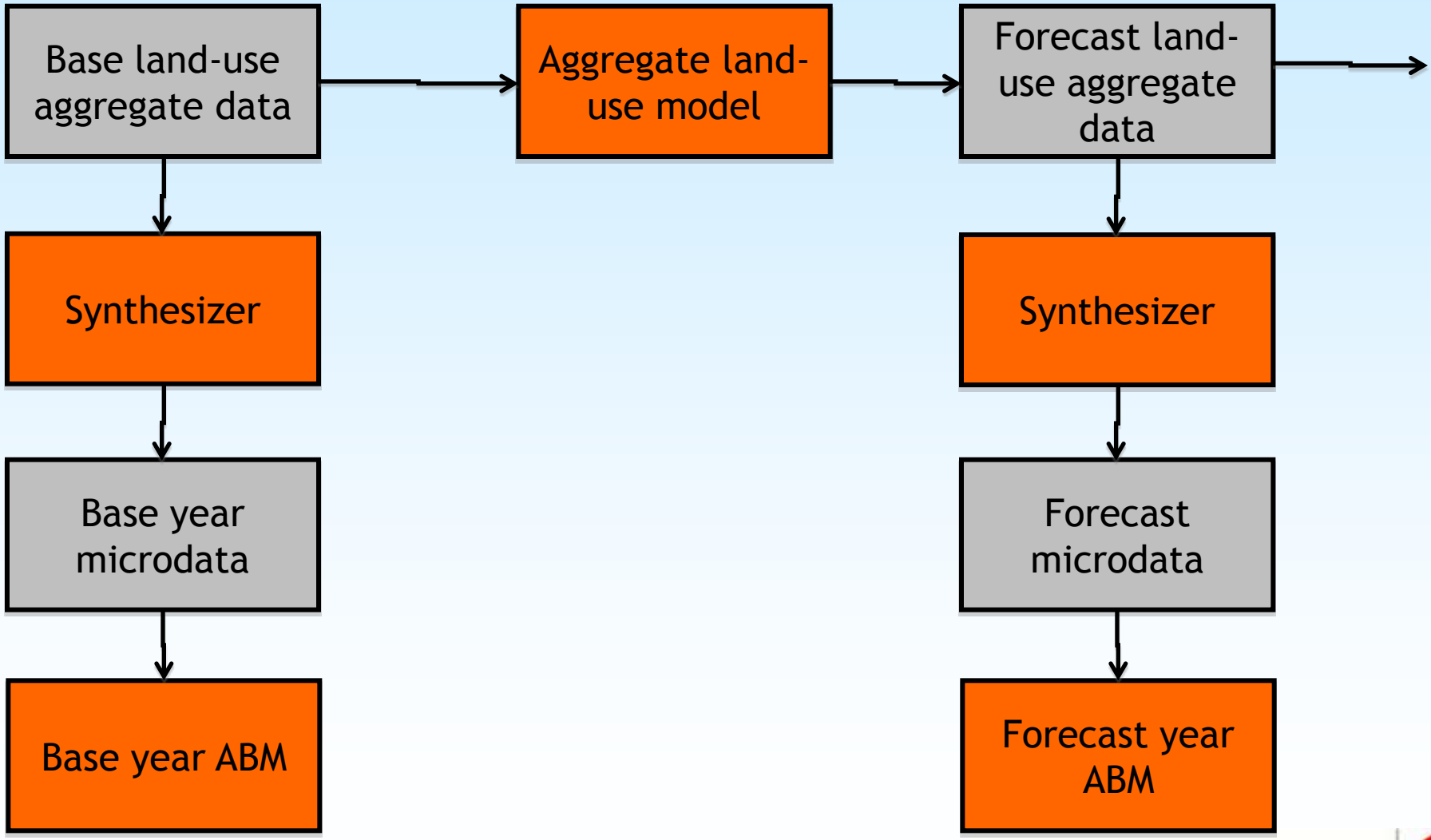
- most or all of these methods are themselves microsimulation methods with a random (Monte Carlo) component – as a result and one population is one "draw" from a large or very large number of populations that could be drawn
- the survey that is expanded by the synthesis may not provide all of the variables needed, or they may not make sense after the expansion.

Synthesizing forecast year data

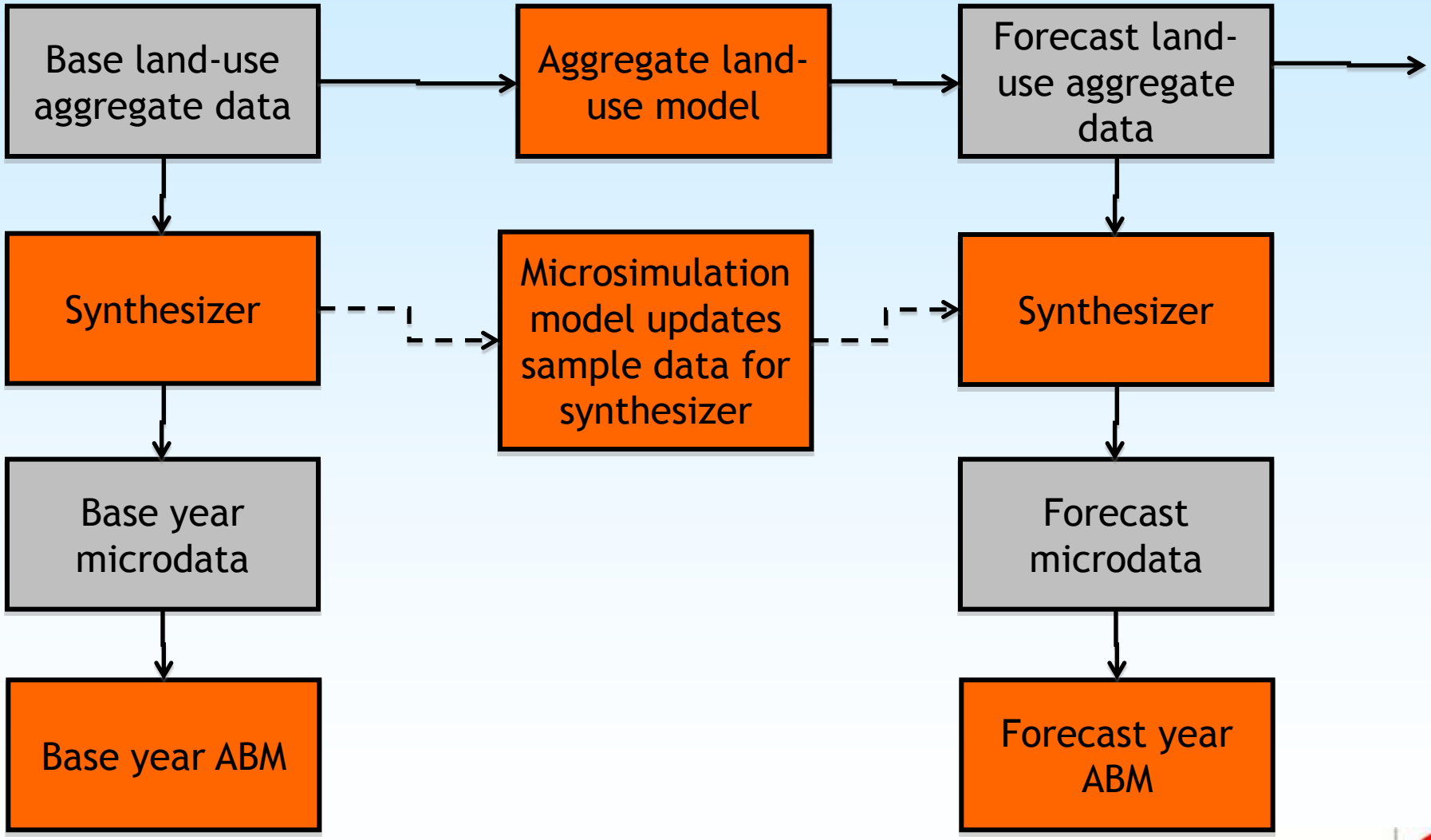
Two broad approaches:

- "conventional" – an aggregate method is used to produce household and/or population forecasts at a zonal level, and a synthesizer process (similar to that used in the base year) is used to produce the household/population data needed for the activity-based model
- "alternative" – a microsimulation model is used which takes the base year synthetic population as input and forecasts how that population changes over time so as to arrive at the forecast year population .

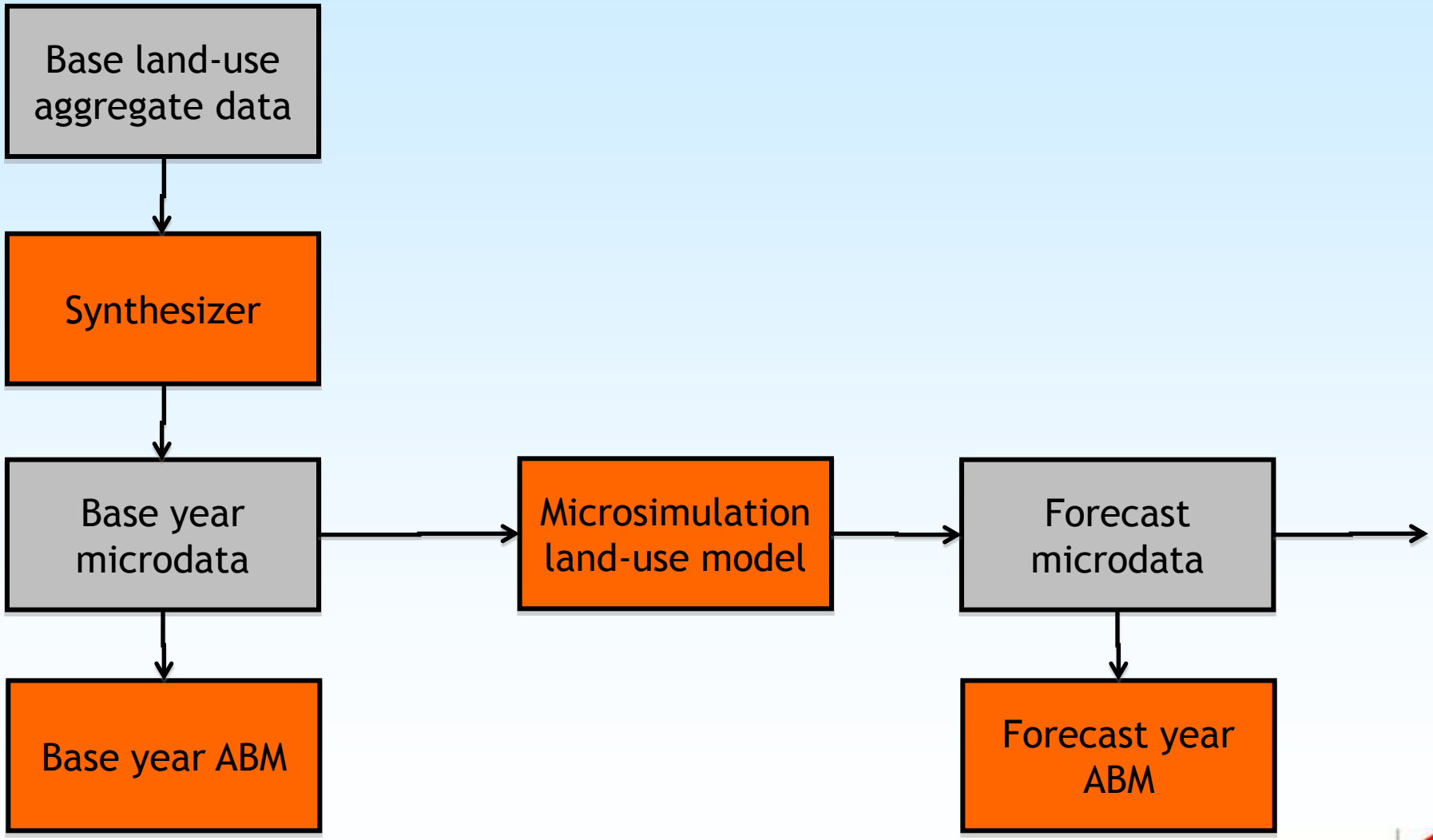
Conventional approach



Conventional approach + “evolution”



Alternative approach



Advantages of the conventional approach

- Doesn't require a spatially detailed microsimulation land-use model (and can make use of an existing aggregate land-use model if one is available)
- If the aggregate forecasts are for the same variables at the same spatial levels as the base constraints, may use exactly the same synthesizer as the base year process
- Doesn't matter how the aggregate land-use forecasts are produced
- in particular, this approach can be used where the aggregate land-use forecasts are subject to an element of adjustment to reflect political aspirations
- The aggregate land-use forecasts are not subject to stochastic error

Advantages of the alternative approach

- Should achieve higher consistency over time (eg in ages) – more important when modelling short time steps or finer categories
- Can forecast **all** the required variables from a dynamic process taking account of changes in relationships between variables (eg cohort effects) and drawing on our understanding of change over time (difficult to control everything in one aggregate model)
- Can work at whatever level of spatial detail is required (which may allow much more detailed representation of land-use policies)
- Can exploit the other advantages of microsimulation land-use modelling eg
  - not constrained to working with a limited number of household/person categories
  - rule-based rather than trade-off calculations.

Choice of approach – a suggested answer

- If the context is to produce a single land-use scenario (or a very small number of alternative land-use scenarios) purely as the basis/bases for testing transport interventions, then
  - a dynamic land-use microsimulation would not be a justifiable investment
  - if the scenarios are to be produced or modified by consultation, the conventional approach is the only possibility.
- If the objective is to look at the **local** effects of **specific** land-use or transport planning decisions, it may be preferable to use an aggregate land-use model because in a microsimulation land-use model the results of interest would be lost in stochastic variation
- If the objective is to look at the **regional** effects of **regional** land-use and/or transport policies (many specific localised decisions) a microsimulation land-use model may be appropriate.

Stochastic variation – the problem

- Microsimulation models that involve Monte Carlo simulation have the property that each set of outputs represents one random "draw" from an extremely large population of possible outputs
- In principle this is desirable because we could run the model repeatedly, look at the distribution of results and reach conclusions about the uncertainty of the forecasts
- In practice it is more often a problem than an advantage
- A particular problem is that we can get noise in the results caused by artificial features of the modelling process rather than by the "real randomness" of the processes we are trying to represent
- Note that the "real randomness" includes effects of the order of households and persons in the sample

Stochastic variation – a treatment

We have (re)invented a “noise reduction” process in which

- we create tables of random values, which can be chosen by the user for each run of the model;
- each use of a random value in the model deterministically generates a integer which is used to look up the corresponding value in one of the tables;
- for “normal”, “noisy” runs we run the model using a different table for each run;
- for “noise reduction” tests we run the model repeatedly using the same table for each run.

This means that in comparing “noise reduction” tests run using the same table, differences arise only from real differences in the factors affecting the modelled choices.

Conclusion

- There are different ways of generating and/or forecasting synthetic populations – appropriate methods depends on resources and on the requirements of the project
- Stochastic variation in the outputs of microsimulation models is a major practical issue for using them
- Our discussions about population modelling shouldn't completely distract us from modelling employment and the impacts of spatial policies on economic performance!