Abstract: This paper discusses the latest enhancements in the microsimulation package SimDELTA, which has been developed by David Simmonds Consultancy. SimDELTA combines a microsimulation model of individual and household changes with the aggregate model DELTA, which forecast household migration between different areas, the location of investment, the pattern of production and trade, the location of jobs, floorspace development and redevelopment, area quality and other components and usually linked to an appropriate transport model.

Keywords: Urban models, microsimulation, stochastic variation, household location, car ownership

1. INTRODUCTION

The SimDELTA package is a variant version of the DELTA package, the variation being that SimDELTA uses microsimulation rather than aggregate methods to represent households and household members, to model the processes affecting them and to forecast the choices they make. DELTA itself is a package for land-use/economic modelling, generally used in interaction with a transport model to build land-use/transport interaction (LUTI) models. The package is based on sub-models representing different processes of change, with most of the interactions between these processes being gradual rather than simultaneous. The logical linkages between the different components are much more complex, with many time-lagged relationships and, of course, many feedbacks both positive and negative. The development of DELTA started in 1995. Its design and applications have been widely reported including several papers to CUPUM (Simmonds and Feldman, 2005, on the design and modelling issues; Bosredon et al (2009) and Feldman et al (2009) on applications). The DELTA process-oriented structure
facilitates the introduction of alternative models such as the use of SimDELTA for the household and individual changes.

SimDELTA is mainly intended to forecast the future number of households living in each zone of the study area and models the following processes:

- Individual demographics and other changes: ageing, survival, birth/multiple birth, entering, leaving, re-entering or retiring from the labour market, educational status, acquiring driving licence, becoming permanently sick, moving to institution.
- Household changes: separation, couple formation, marriage, absence from households, student households and other shares, obtaining/losing car, household income, housing affordability.
- Household location: seeking to move, housing tenure choice, dwelling choice, housing prices or rents, location choice, location/relocation, in/out migration.
- Employment: seeking to change job, job and workplace choices, wages, accepting/rejecting candidate, and accepting/rejecting job.

The model takes into account the impact on households of changes in the transport system and hence in accessibility, of changes in the supply and characteristics of housing and of changes in the number and distribution of workplaces. Households and their members go through a series of processes associated with the medium- and long-term responses to all these (see Feldman et al., 2007a,b for more detail).

Rather than creating a wholly new application, the decision was taken to base the new model on the existing South and West Yorkshire Strategic Model, SWYSM (see Simmonds and Skinner, 2004). The starting point for the development of the microsimulation elements was the MASTER (Micro-Analytical Simulation of Transport, Employment and Residence) model, which was previously developed at University College London (Mackett, 1990, 1992, 1993). The resulting application to the South and West Yorkshire model was called SWYSim, and it covers 5 large areas covering virtually all of the South and West Yorkshire areas and some adjoining territory; these areas are represented at ward level, giving 283 microsimulation zones. The rest of the SWYSM modelled area forms a set of 18 external zones for the SWYSim application. These cover the rest of the South and West Yorkshire plus the larger adjoining areas of Greater Manchester, Humberside, Lincoln, Nottingham, Derby, Stoke, East Lancashire and Yorkshire. In these zones population is still modelled in standard DELTA, i.e. at the aggregate rather than microsimulation level.

The main data sources used in the present study were the UK Census of Population, the 1991 UK Samples of Anonymised Records (SARs), and the British Household Panel Survey. The base year in the model is 1991 and the initial synthetic databases were created for 1991. The preparation of the initial synthetic population is a major microsimulation exercise in itself; it was carried out by a team at the University of Leeds School of Geography and described in an earlier CUPUM paper (Feldman et al, 2005).

---

1 We used the 1% sample of households and individuals in those households extracted from the 1991 Census of Population. Unfortunately, the 2001 Household Sample of Anonymised Records was not yet released when the project was carried out.
This paper focuses on two areas which have been the subject of recent work: the issues arising from the validation of the model and more specifically with the assessment of stochastic variation, and the micro-modelling of car ownership. Although stochastic variation is considered to be one of the key weaknesses of microsimulation models, its impacts and implications have not been extensively investigated. The analysis of stochastic variation contains two main parts: first, the impact of the population sample size on stochastic variation is assessed; then a noise reduction function is applied in order to eliminate the impact of noise (stochastic variation) on a policy testing.

Further enhancements in the microsimulation part of the SimDELTA model have been introduced in order to reduce its dependence on the aggregate model (DELTA), hence to increase its self-containment. The new car ownership module simulates car ownership choices at individual level based on socio-economic characteristics. The derived results are validated against real data and compared to the forecasts of the aggregate car ownership model.

2. STOCHASTIC VARIATION – ISSUES AND TREATMENTS

Many of the components in SimDELTA use Monte Carlo microsimulation of choice processes. In each such choice process, the probability of each outcome is calculated for the household or person in question, and then a random number value is used to determine which outcome is forecast to arise. If the random values are genuinely random, then each forecast produced by the model will to some extent be different, even when the all of the inputs are identical. In a constrained system such as a land-use model, where the location of households is constrained by the availability of housing and the availability of housing is constrained by planning policies, the scope for the aggregate outputs (total households by zone) to differ will be limited, but there is far more scope for the more detailed results (for example persons by age and zone) to differ significantly.

In principle this is a desirable property of the model, because it can show the uncertainty associated with the results – though to do so requires the model to be run repeatedly, in order to build up the distribution of the results, where a deterministic model would only need to be run once; this can be a serious practical problem. Within this, however, there is a distinction to be made between “real randomness” and what we regard as “noise” caused by the interaction between the random number processes and the way in which the model is operated.

A key point to note about the random number processes is that in practice microsimulation models work with pseudo-random number sequences which start from a controllable “seed” value. For any one seed value, the series of random numbers generated will be the same. Each Monte Carlo choice made takes the next value out of this series. The fact that the sequence is controlled by the seed value is helpful in testing model software, in that it can allow a model run to be repeated with exactly the same sequence of random values, for example to test whether a specific bug has been successfully corrected.

The problem that still arises, even when the random sequence is started from the same seed, is that the random values used for any one specific choice will only be the same in two tests if that choice comes at exactly the same point in the series of random numbers used. The element which we consider “noise” is that where, for one random number seed
that the outcome of one decision or other process, for the same household or person, in the same year and in identical circumstances, is different in two tests because an unrelated change elsewhere in the model has used more, or fewer, random values, and hence causes a different random number to be used in the decision or process of interest. This “noise” is an artefact of the way in which the random numbers are produced and used within the software. It is both a nuisance in checking software or other model changes, because it makes it very difficult to ensure that each choice uses the same random value, and in the comparison of model runs where we may wish to assess the effect of different inputs (for example, different planning policies) whilst keeping all other effects – including random ones – constant.

The method we have adopted to deal with this – when appropriate – is essentially to use pre-prepared look-up tables of random values rather than the pseudo-random number sequence. We create a number of files each containing a large set of random numbers drawn from a uniform distribution in the range 0 to 1, using a different seed for each file. Any group of tests for which we want to control the random variation should use one of these files (chosen by the user). Each Monte Carlo Simulation (MCS) step within the model takes its “random” value from a deterministically-defined position in the chosen table. This means that a specific choice or process for a specific household or person, in a specific year, will always use the same cell of the chosen table, and hence the random value for that specific choice or process will change only if the user changes the table used. The deterministic position in the look-up table is calculated as a function of variables identifying the process being made and the household or person affected, e.g. it uses the nth cell of the matrix where n is obtained from (household_ID)*(year)*(arbitrary value for that particular choice or change). Similarly each MCS step for a person outcome should be based on $n = (person\ ID)*(year)*(arbitrary\ value\ for\ that\ step)$. The arbitrary values are constants in the code and have different values for each process i.e. one constant for the mortality process, one for the fertility process, and so on. If the calculated value of $n$ is greater than the length of the look-up table, it is truncated to a value within the table, e.g. if $n$ is calculated as 2568974 but the look-up table contains only one million values, the process uses the 568974th value.

This noise-reduction (or noise-elimination) treatment allows us to control the Monte Carlo processes so that in testing modifications to the model, we can be sure that differences were due solely to the modifications, and in testing policies interventions, we can be sure that differences between tests were due solely to the policies introduced (again, with the same contrast to use of a pseudo-random sequence). Specifically, it eliminates the “noisy” variation in results which occurs when a small change to the model design or inputs results in one household using one more, or one less, random value; in the conventional sequence-based process, this will cause every subsequent household in the model to use a different random value for every choice. With the table-based procedure, unrelated choices can be controlled to use the same values, and arbitrary variation linked to the order in which households are processed can be avoided.

3. ASSESSMENT OF STOCHASTIC VARIATION IN SWYSIM

The evaluation of stochastic variation is based on the results of different runs of the model using different random seeds. More specifically, SWYSim was run using a 10% sample of
the population 10 times – using 10 different random seeds – over a period of 5 years. Annual population changes were estimated by taking the average of the results of the 10 runs. The results for the period 1991-1996 are presented in Figure 1. It can be seen that most of the population changes occur in the same areas every year, i.e. over this lustrum there are areas of continuing change and areas of comparative stability.

![Figure 1: Annual population changes](image)

![Figure 2: Absolute change of district population 1991-1996; comparison of 10 runs using different random seeds; districts are ordered by size](image)
Figure 3: Percentage change of district population 1991-1996; comparison of 10 runs using different random seeds; districts are ordered by size.

Figure 2 and Figure 3 show the results of the 10 different runs aggregated to the district level. The districts are presented in the ascending order according to their population in 1991. The figures indicate that there is, as one would expect, a tendency towards a greater absolute variation in population change in large districts, and towards a greater proportional variation in smaller districts. A similar tendency is seen at the ward level² (Figure 4 and Figure 5).

Figure 4: Annual absolute population changes by ward, 1991-1996, comparison of 10 runs using different random seeds; wards are ordered by size.

² There are a total of 8850 wards in England and Wales with a minimum size of 100 residents or 40 households. 283 wards cover the SWYSim area.
Figure 5: Annual percentage population changes by ward, 1991-1996, comparison of 10 runs using different random seeds; wards are ordered by size.

Figure 6 shows the relative standard deviations of population changes by SWYSim area from 1991 to 1996 calculated from 10 runs and population changes weighted by the relative standard deviations. It can be seen that the relative standard deviation remains high in most cases with the exception of wards where the biggest population changes occur (see Figure 1). The results on population changes and standard deviation are combined into an index ranging from -3 (the highest negative changes) to 3 (the highest positive changes) as presented in Table 1 in order to identify the wards where changes are more likely.

Table 1: Index of population changes\(^3\)

<table>
<thead>
<tr>
<th>Index</th>
<th>Population Change</th>
<th>RSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>&gt;50</td>
<td>&lt;= 10</td>
</tr>
<tr>
<td>2</td>
<td>&gt;50</td>
<td>&gt;10 &amp; &lt;=20</td>
</tr>
<tr>
<td>1</td>
<td>&gt;50</td>
<td>&gt;20</td>
</tr>
<tr>
<td>0</td>
<td>&lt;50 &amp; &gt; -50</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>&lt;= -50</td>
<td>&gt;20</td>
</tr>
<tr>
<td>-2</td>
<td>&lt;= -50</td>
<td>&gt;10 &amp; &lt;=20</td>
</tr>
<tr>
<td>-3</td>
<td>&lt;= -50</td>
<td>&lt;= 10</td>
</tr>
</tbody>
</table>

\(^3\) the limits were determined after some tests but they are not theoretically justified.
4. NOISE REDUCTION IN SWYSIM

In order to assess the performance of the SimDELTA noise reduction function the following test has been executed in SWYSim. The model was run for the 10% sample of the population, using the original and a modified household databases and different modes (on and off) of the noise reduction function. The order of households in the database was reversed so as to analyse whether the noise reduction function assigns the same probability to each process executed by a household or a person at a specific time independently of the order of households and persons in the database. Using the noise reduction function, the same random number is chosen for each household to simulate a specific process in a specific year independently on a position of this household in the list. Alternatively, with the noise reduction function turned off, a random number assigned to each household is dependent on the order of households in the database: different random numbers will be chosen for the same household when simulating a specific process in a specific year in each run.

The noise reduction tests were carried out with all households in the model and with a 10% sample. The following two graphs (Figure 7 for a 10% sample and Figure 8 for the whole population) illustrate percentage differences between the population forecasts in 1993 by wards produced by two consecutive runs of the model. The two series in each scatter plot refer to the different modes of the noise reduction function (on and off). The results indicate first that, as expected, the level of stochastic variation is greater when the model is run on a 10% sample of households than when it is run on 100% data. The proposed noise reduction method, referring to a deterministically-chosen cell of a predefined random number table rather than to the next value in a pseudo-random sequence, significantly reduced the effect of reordering the data in the 10% sample run and largely eliminates the effect in the 100% sample run.
Figure 7: Percentage difference of total population between the two runs by ward (the model was run on a 10% sample of the population). The wards on the x axis are ordered from the smallest to the largest.

Figure 8: Percentage difference of total population between the two runs by ward (the model was run on a 10% sample of the population). The wards on the x axis are ordered from the smallest to the largest.
5. CAR OWNERSHIP MODEL

The SWYSM land use model has a car ownership sub-model which works entirely in terms of the zonal probabilities of a household of a particular type owning no car, one car or two-plus cars. In the original version of SWYSim these probabilities were input to the microsimulation model and used to generate the probabilities for individual households of acquiring an additional car or giving up a car. Recently, a new car ownership component of the SimDELTA model, based on Hanly and Dargay (2000), was developed. The Hanly and Dargay’s model was estimated using an ordered probit specification on British Household Panel Survey (BHPS) data for the period 1993-1996. The model estimates impacts of the past car ownership on the current car ownership by using lagged endogenous variables such as the number of cars owned by the household in the previous period and lagged dummies for different levels (0, 1, 2 and 3 cars).

The mathematical specification of the model is:

\[
Y^*(i, t) = X(i, t) + \gamma Y(i, t - 1) + \varepsilon(i, t)
\]

\[
Y(i, t) = \begin{cases} 
0 & \text{if } Y^*(i, t) \leq 0 \\
1 & \text{if } 0 < Y^*(i, t) \leq \mu_1 \\
2 & \text{if } \mu_1 < Y^*(i, t) \leq \mu_2 \\
3 & \text{if } \mu_2 \leq Y^*(i, t)
\end{cases}
\]

where \(Y^*(i, t)\) is the latent ordinal preference index for individual \(i\) in time period \(t\);
\(Y(i, t)\) is the observed ordered-response choice (number of cars owned);
\(X(i, t)\) is the vector of exogenous explanatory variables (adults (older than 17 years old), number of employed persons, number of children, pensioner household dummy, population density, 1, 2 and 3 car dummies (correspond to car ownership in the previous year), and attrition);
\(Y(i, t - 1)\) is a state variable (car ownership in the previous period);
\(\beta\) is the vector of parameters;
\(\gamma\) is the weight for the state in the previous period; and
\(\mu_1, \mu_2, \mu_3\) are constant thresholds, and
\(\varepsilon(i, t)\) is the random variable \(E[\varepsilon(i, t)] = 0\).

The parameters used in SimDELTA are those estimated by Hanly and Dargay (2000), and are shown in Table 2.
Table 2: Estimated parameters of the car ownership model (Hanly and Dargay, 2000)

| Parameters                                      | Coeff | P[Z>|z] |
|-------------------------------------------------|-------|---------|
| Income                                          | 0.010 | 0.000   |
| Adults (older than 17 years old)                | 0.325 | 0.000   |
| Number of employed persons                      | 0.150 | 0.000   |
| Number of children                              | 0.002 | 0.878   |
| Pensioner household dummy                       | -0.259| 0.000   |
| Population density                              | -0.005| 0.000   |
| 1 Car dummy (last year’s car own.)             | 2.507 | 0.000   |
| 2 Car dummy (last year’s car own.)             | 4.040 | 0.000   |
| 3 Car dummy (last year’s car own.)             | 5.008 | 0.000   |
| Attrition variable                              | -0.347| 0.021   |
| Constant                                        | -1.304|         |
| $\mu_1$                                         | 2.802 |         |
| $\mu_2$                                         | 4.781 |         |

The key processes of the microsimulation car ownership model are the following. For each household the latent ordinal preference index in the current period is estimated. The household and person databases of SWYSim provide the data for all exogenous explanatory variables except the attrition variable, which has to date been treated as a constant. The values of the preference index are compared to the $\mu$ values (Table 2) in order to determine the “preferred” car ownership level for the household. If the current car ownership level for the household is different from the “preferred” level, then a change in car ownership towards the preferred level may occur; this is controlled by a Monte Carlo process. Only one car may be added to or subtracted from the household, so in any one year the car ownership level of each household can change only by one unit.

A final step is needed to moderate the adjustment from the initial levels of car ownership to the “preferred” values calculated in the model; without this a rather abrupt change in car ownership may occur in the first modelled year as the model goes from the observed base year values to the (deterministic) forecast values. The moderation method used is itself a Monte Carlo simulation, with the probability of a household changing car-ownership being related to how well the original Hanly and Dargay model fitted the data in terms of predicting the proportion of households in the target category; in other words, the more certain the original model was in predicting a particular level of car ownership, the more

---

4 On more recent reflection, treating the attrition variable as a constant is certainly not correct, but it is difficult to improve on this in the present case. From descriptions of the Heckman two-stage method, we know that the attrition variable used by Hanly and Dargay would have been the output, for each household, of a preliminary model calibrated to estimate the probability of a household dropping out of the sample (the sign on the coefficient tends to confirm this). Without the details of that preliminary model it is impossible to recreate the attrition variable within the microsimulation. Those details were not included in the paper on the car ownership model and seem to have been lost since the original work was done; it is not even clear whether the attrition model was calibrated specifically as part of the car-ownership study or was taken from other work on the BHPS – since the problem of attrition is a general one in the use of that and similar panel studies and not by any means unique to the car-ownership question.
likely it is in the present application that the household will adjust its car ownership in that direction. A further constraint, that households only adjust their car ownership up or down by one car each year, is also imposed.

With further resources — including direct access to the BHPS data, which the present authors did not have — it might be possible to calibrate a model that would directly forecast changes in car ownership, i.e. one where the \( \mu \) values would be the thresholds for acquiring or giving up a car (or two cars); this might ultimately be more appropriate to the workings of the model.

5.2 Results

SWYSim was run from 1991 to 2001 using the new microsimulation car ownership model. Work is continuing on the analysis of the results. The main analysis so far has been to consider the elasticity of car-ownership with respect to a change in income. Figure 9 shows the forecast growth in the total stock of cars owned (or more precisely, available for the use of) households within the modelled area. The “Base Case” line shows the growth in this variable with the standard income trend; the “Test” shows the slower growth as a result of imposing an arbitrary 10% reduction of incomes in 1992. It can be seen that the impact of lower incomes on the numbers of cars run by households increases over time.

![Total car fleet, base case and income test](image)

Figure 9: Total car fleet forecasts in base case and income test (income test is a 10% reduction from 1992 onwards).

The implied elasticity of car ownership is shown in Figure 10 measured as the proportional change in the fraction of household owning cars divided by the proportional change in...
income. The elasticity starts very low, increases over the next three years, dips for several years and then increases again. Initial review suggests that the elasticities are similar to those found by Hanly and Dargay, but further work is needed (including checking of definitions) to confirm or refute this. What is evident is that although the car ownership model (as defined earlier) is in itself relatively simple, when applied and combined with the other features of the model it gives rise to seemingly complex dynamics of car ownership changes over time.

![Figure 10: Elasticity of car-ownership with respect to income (from income test: 10% reduction from 1992 onwards).](image)

**CONCLUSIONS**

This paper reports on two enhancements of a microsimulation model SimDELTA, namely, the use of a noise reduction function in the model, and the development of a car ownership model which ensures greater consistency between the population and households microdata and car ownership as well as being more appropriate to the overall structure of the model. We discuss the scale of stochastic variation in a SimDELTA application to the South and West Yorkshire area (SWYSim model), and the impact of a particular method to eliminate arbitrary noise caused by the interaction of sequential Monte Carlo processing of households with a given sequence of pseudo-random numbers. This topic is important in the use of Monte Carlo microsimulation models for
forecasting and (potentially) for policy analysis, and we believe that this method may be of a wider interest beyond the SimDELTA application.

6. NOTES AND ACKNOWLEDGEMENTS

The authors are grateful to the UK Department for Transport for supporting the original SimDELTA project and for permission to continue with subsequent development. Dr Christodoulou’s work on SimDELTA was carried out under a Knowledge Exchange Associateship sponsored by University College London. The opinions expressed here are those of the authors and not necessarily those of the client or sponsoring organizations. The authors are grateful to colleagues at DSC and elsewhere for all their contributions to the projects described here; they remain responsible for any errors of commission or omission.

7. REFERENCES


Feldman, O, Simmonds, D, Zachariadis, V, Richmond, E, and Mackett, R (2007a): SIMDELTA - A microsimulation approach to household location modelling. Paper presented to the 11th World Conference on Transport Research (WCTR), San Francisco, USA.*


Papers marked * are available on the Publications page at www.davidsimmonds.com

