Chapter 3 Review of Microsimulation

3.1. Introduction

We begin the chapter by setting microsimulation within the wider context of the models used to forecast spatial populations of individuals and households. We need to choose an approach from the large number that has been experimented with by researchers in the past fifty years.

3.1.1. A Framework for Forecasting Households

One of the key determinants of domestic water demand is the residential population. As discussed in Chapter 2, the residential population is the base for implementing scenarios of policy impacts, climate change, socio-economic and technological trends, so that all types of domestic water demand models incorporate population in either macro or micro forms. Many dimensions can be employed to classify population estimation and projection method, for instance, macro and micro approaches. In this research we argue that another dimension, that is housing, should be added, since migration of population may be attracted by housing capacity, especially at small geography scale and in a developed nation such as the United Kingdom. Microsimulation could model other entities such as houses, customers or factories, but in this thesis we are focusing at residential population. Thus, we can propose a cross tabulated table (Table 3.1) for classifying population projection approaches with scale (macro/micro) as the row dimension and, principal driver (housing led/population led) as the column dimension. The contents of the table will now be briefly described.

Table 3.1 Approaches to Forecasting the Households in Residential Demand Models

	Principal driver of the projection	
Scale of unit used in model	***Population-led models***	***Housing-led models***
Macro-population/household (i.e. population and household numbers for areas)	Cohort-component model with added conversion of individuals into households	Roll forward of existing housing with demolitions and new housing added. Conversion into households and individuals based on updated housing type and size information.
Micro-population/household (household and individual records)	Dynamic microsimulation model using cohort-component processes and modelling households and individuals together	Microsimulation of households linked to the changes of housing, from a macro model or direct data source.

3.1.2. Cohort-Component Method (Population-Led)

The cohort-component method is widely applied in population projection, which is based on past population structure by age and sex. These cohorts of population change through the fundamental population processes of birth, death, and migration. To predict future population, we forecast the future trajectories of the component rates. Multiregional cohort-component models were devised by Rogers (1968, 1975, 1985, 1995) and by Rees and Wilson (1977) and Rees (1984). These models disaggregate the population further by the variable of space. Therefore, changes of the populations in sub-regions include their own birth, death components and the migrants from and to other sub-regions. In addition, international migration into regions is usually treated as a net figure. The difficulty in these applications lies in the estimations of the age, sex and region specific intensities for the components in the projection years. These models are based mainly on the average behaviour of groups of individuals. However, researchers such as Rees (1986a) recognise that families and households are the basic units making the decisions on processes such as migration, marriage and childbirth. Duley (1989) suggests that it is more appropriate to take the household context, for households are the basic unit of decision making, production, distribution and consumption in many cases. Further, Clarke (1986) outlines how household projections are important inputs to a range of planning and policy issues that population projections are unable to handle, for

example, demand in housing markets, the location of retail facilities and personal and health care services. As in individual based population models, macro household population projections set up household population structure as types of households, by variables such as age of household head, household size and accommodation type (Parsons *et al.,* 2007) or household composition (Brouwer, 1988, Ermisch and Overton, 1985, Keilman, 1988). Estimation of the transition probabilities are the main complexities for the categories of households, part of which is caused by lack of data.

3.1.3. Dynamic Microsimulation Models (Population-Led)

Dynamic microsimulation models usually employ the cohort-component approach to provide the probabilities of surviving and giving births for modelling the demographic aspect of the model (van Imhoff and Post 1998), which will in turn result in composition changes of household units. To simulate the death and birth events in a microsimulation model, target persons need to be selected from eligible candidates using either a discrete or continuous time frame simulation process, which will be examined in section 3.2.2 in detail. In both approaches demographic probabilities are the essential inputs, which are usually collected from published statistics which provide cohort-component information. Cohort-component approaches usually disaggregate the population by age and sex only, the limitation of which will be naturally inherited by the microsimulation modelling and so groups of the micro population (by age and sex) have to share the same probabilities. The situation may be helped by auxiliary data, such as social class differential mortalities or racial weights for fertilities from Panel or Longitudinal data, to increase the precision of the modelling. However the average assumption will not be improved much by the auxiliary data unless the panel data has a great coverage so that the disaggregated probabilities can be more accurate and representative.

3.1.4. Macro Housing-Led Models

Many researchers have accepted and implemented the concept of housing-led population dynamics. For example, Parsons *et al.* (2007) project macro household populations in the Thames Gateway relying on the housing forecasts and strategies from regional and local authority planning reports. Similarly the projection in Hollis (2008)

adds past and projected changes of homes to the 2001 London household population at borough level.

3.1.5. Microsimulation Models with Housing Variables

A micro example of the housing led approach is the model of Wu *et al.* (2008). The model simulates flows of migrants in Leeds using a spatial interaction model involving housing and the model also simulates moves of university students through an multi-agent model based on the vacancy of renting spaces and the concentration of the same types of students (first year undergraduate, other year undergraduate, master and doctorate students).

Van Imhoff and Post (1998) provide a comprehensive comparative review of macro and micro simulation, which we summarize here. Microsimulation and macrosimulation are two alternative methods for making similar statements about the future. Both approaches are based on models, need assumptions about the future paths of their input variables and must always contain the time element. Further, when the number of variables in the disaggregation of the macro population increases, macrosimulation require a very large array of transitions between states whereas a microsimulation model requires storage proportional to the population size. Moreover, the number of probabilities in a macromodel expands with the square of the number of states. The number of transit probabilities will become unmanageable in macro models that handle a large number of states. Since explanatory variables are presented in the data structure of a list of micro units with attributes, complicated relationships between variables can be handled well. Microsimulation models using households as micro units are modelled at the decision making level (the couple, the household) of many events such as marriage, migration or divorce. This is difficult to do in the individual based macro population models. Finally, microsimulation models can provide much richer results, such as variable combinations not otherwise available, and so higher precision.

Based on the brief review of approaches to forecasting households outlined in Table 3.1 we argue that a microsimulation approach will be superior to a macrosimulation approach. Thus, for deciding specific techniques to be employed for this thesis, a review of the state of art of microsimulation modelling will be presented in section 3.2. The

chapter introduces spatial data reconstruction methods in section 3.3. It provides a snapshot of selected microsimulation models in section 3.4 and describes a prototype microsimulation model in section 3.5. The chapter concludes by examining the lessons learnt from the literature and the development of the prototype model and, also designs the structure of spatial water demand microsimulation model in section 3.6.

3.2. Microsimulation Modelling

A microsimulation model is a convincing recreation of a set of conditions or real-life events (Chambers 2005) in order to analyse the behaviour of a system based on very small entities, which could be atomic particles, artificial species, firms or individuals, families and households (Spielauer and Vencatasawmy 2001). It is argued that the term "microsimulation" commonly referred to is data-based or data-driven microsimulation. Empirical data and statistics are used to forecast the behaviour of an empirical population, whereas agent-based simulation works on the basis of rules and 'intelligent' behaviour and is designed to study the dynamics and patterns of artificial societies which result from the interactions of artificial species that follow certain rules (Spielauer and Vencatasawmy 2001). It is recognised by social scientists that Orcutt (1957) was the first advocate of microsimulation in economics, in reaction to the shortage of detailed macro economic and demographic models for policy analysis (Holm *et al.* 2003). Orcutt (1957) suggested that aggregate socio-economic models are limited in their ability to predict the effects of alternative governmental actions, or to test hypotheses and to estimate relations, and fail to predict distributions of model units in single or multi-variable classifications. A microsimulation model takes micro level units as the basic units of analysis to investigate the effects of social and economic policies (O'Donoghue 2001). In other words, most microsimulation models are developed as policy 'laboratories', in which individual behaviour and so the distribution impacts of policy changes can be forecast and then investigated.

Microsimulation models allow the heterogeneity of the state of a system to be fully represented and stored by the most basic units. This avoids losing much information which occurs in macro models because of the limited disaggregations that can be made. More specifically, non-linear relationships concealed in these disaggregations, which are untraceable by macro modelling, could be revealed and updated in a

microsimulation model (Orcutt 1957). Achieved knowledge about the behaviour of decision-making units can be incorporated and the output can easily be aggregated to the levels suitable for answering research and applied questions (Holm *et al.* 2003).

Conventionally, microsimulation models may be classified into static and dynamic types, while the subtype of dynamic longitudinal microsimulation can be recognised within the dynamic type (Baldini 2001). A static microsimulation model is representative of the population at a given time (Pudney 1994), which is in the form of a list of micro units with their characteristics, for instance, persons with their characteristics such as age, sex and other social economic variables. The uses of a static microsimulation can be easily explained by the income and tax example. If the micro data for a population with pre-tax and personal socio-economic characteristics are available, experimental tax policies such as tax rates and personal allowances can be applied, and then the effects of these policies on disposable income and the distribution of post-tax income can be predicted and analysed. No characteristics of the members of the micro data are changed during the process except the disposable income. Recent developed static models are able to adjust or create the micro data to fit future aggregate data (Ballas *et al.* 2005a, 2005b, 2005c, 2005d), so that policies can be experimented using the future population as well. Static models are more commonly used to evaluate the immediate effects of economic policies (Baldini 2001). O'Donoghue (2001) recognises that dynamic microsimulation modelling is not only able to update the characteristics of the micro units caused by the stimulation of endogenous factors called behavioural reactions, such as labour supply responses to changes in government policy, but can also project them over time to include demographic processes and social economic transitions, such as ageing, mortality, fertility or social and geographical mobility. Dynamic longitudinal microsimulation is also referred to dynamic cohort microsimulation which simulates the events related to a group of micro units born in the same period from the birth to the death of the last member, and focuses only on a life-cycle analysis (Baldini 2001). The advantage is that the information on the complete individual life-cycle of each member is available.

Spatial or geographical microsimulation adds space or geography as a new dimension to microsimulation models so that the spatial distribution of policy impacts can be obtained. Hence dynamic spatial microsimulation not only models the demographic and

social economic aspects in a micro system but also the geographical or spatial processes such as household location, changing residences due to marriage and divorce. Hägerstrand (1953) pioneered work in spatial microsimulation theory and provided a conceptual framework for the micro level analysis of spatial dynamics based on a representation of actors, resources, and other objects located in a micro level time-space framework (Holm *et al.* 2003). Recent example of spatial microsimulation models include: SYNTHESIS (Birkin and Clarke 1988), UPDATE (Duley 1989), Williamson' elderly care population model (Williamson 1992), SimLeeds (Ballas *et al.* 2005c), SimBritain (Ballas *et al.* 2005a) and SVERIGE (Holm *et al.* 2003). Ballas *et al.* (2005a) argue that spatial microsimulation "provides useful information on socio-economic trends as well as on the possible outcome of policy reforms, at different geographical scales" (p14), so that the models can be used to estimate the geographical impacts of national policies and inform decisions on the revision of these policies on the basis of their likely spatial as well as socio-economic distributional effects.

Basically, three components make up a dynamic microsimulation model: the initial base data, the simulated processes and the policy instruments (O'Donoghue 2001). The same statement could also be applied to some of the static models which have the processes of adjusting or constructing the microdata to fit future aggregate data. One more component, alignment of the model to targets, may be included, since alignment (particularly verification) has been attracting more and more attention.

3.2.1. Data

The methods used to prepare input base micro data with spatial detail for a spatial microsimulation model are examined in the section 3.3. The following is a brief description of the major available population data sets in the United Kingdom.

- **The Samples of Anonymised Records (SARs) from the 1991 Census of Population**: The SARs are released as abstracts of the Census microdata to provide the information for anonymised individuals and households. The 1991 Individual SAR holds 2% of the total population containing 1.1 million individuals and visitors in private households and communal establishments. The 1991 1% household SAR has 216,000 households containing around half a million members. The 1991

Household SAR has all Census variables and derived household and family level variables. The 1991 Individual SAR has identifiers for 278 geographical areas, which are either whole local authorities (LAs) or combinations of local authorities joined to ensure a minimum threshold population of 120,000 people. The geographical identifiers in the 1991 Household SAR refer only to the Registrar General's Standard Regions. The linkages between individuals can be found in the 1991 Household SAR. Individuals belong to a family and households are made up of one or more families or a set of unrelated individuals in a pseudo-family. The family is the middle unit between household and individuals. A family can have two generations at a maximum in which the younger generation must be single (never married) and have no obvious partner including cohabitee or offspring (Office for Population Censuses and Surveys, 1992). Individuals in a pseudo-family do not necessarily have a real family relationship among them, for example, student households. This type of 'family' is given a special family identification of '0'. A household can have multiple families though this is unusual.

- **The 2001 Census SARs.** The second version of the 2001 Individual SAR (CCSR 2005) was released in 2005 with slightly more disaggregation for some variables such as age than in the first release, which resulted in complaints by users about the lack of detail. The 2001 Individual SAR is a 3% sample of the national population of individuals, and numbers about 1.76 million records, which is about 50% larger than 1991 Individual SAR. However, the amount of detail in the 2001 Individual SAR is much lower than in the 1991 Individual SAR. The geography used is Government Office Region (in England) rather than LAs or LA combinations in 1991. Ages are grouped from 16 to 74 into seven 5-year bands. Only 17 industry categories are reported compared with 42 in 1991 and there are only 16 country of birth categories compared with 42. The 2001 Household SAR (1%) was released in 2005 for England and Wales comprising about 200 thousand households and circa 500 thousand individuals. Less detail compared to its individual counterpart was produced and to its 1991 equivalent. For example, no sub-national geography was included.

- **The ONS Longitudinal Study (LS) (Creeser *et al.* 2002):** A sample of about 500,000 records (selected by using 4 birthdays out of 365 or just over 1% of the

population) records from the Censuses of 1971, 1981, 1991 and 2001 have been selected and linked at individual scale in a LS database. The database holds full Census information without confidential protection measures applied for these records, although there is strict control of what processed data can be extracted or which analyses can be done. Not only are the individual records themselves in the database but also all members of their households. Vital events such as births on the four LS birthdates (which are kept secret), immigration of people with one of those birthdates, deaths, births and infant deaths have been linked to the LS individuals. The data can only be analysed in a safeguarded 'setting' (at the Office for National Statistics, through the Centre for Longitudinal Study Information and User Support or CeLSIUS). A variety of analyses can be carried out ranging from aggregation of the micro units to testing statistical models using software packages. The coverage of LS can be utilised to complement other smaller but specialised microdata sets such as the British Birth Cohorts or more frequent longitudinal surveys using shorter time interval such as British Household Panel Survey (BHPS) (Openshaw 1995).

- **Census Area Statistics (CAS) and Standard Tables (ST):** Small Area Statistics (SAS) have 86 tables consisting of about 9000 counts, of which the counterpart in 2001 Census, called Census Area Statistics (CAS), have fewer counts in simpler tables. Area data tables having more disaggregated counts and tables than the SAS are called Local Base Statistics (LBS) in 1991 Census and Standard Tables (ST) in the 2001 Census. The LBS contains 20,000 counts in 99 tables. The major difference between SAS and CAS on the one hand, or LBS and standard Tables on the other hand is that the former are slightly more aggregated and available for lower geographies. LBS and Standard Tables are only available from higher geography to ward level, whereas CAS and SAS are also published at ED (Enumeration District) and OA (Output Area) level respectively. The figure that follows shows the number of tables and average counts per table of the products from 2001 Census.

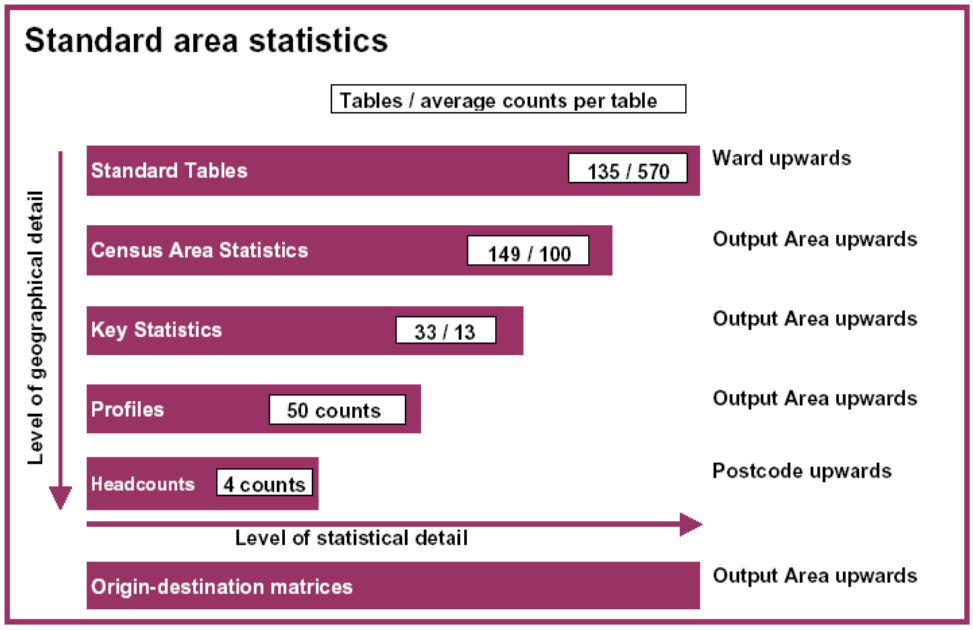

Figure 3.1 The Standard Area Statistics of the 2001 Census, sourced from ONS (2004b)

The relationship between 1991 and 2001 Census area statistics data sets is illustrated by the Table 3.2.

Table 3.2 The Area Statistics from the 1991 and 2001 Censuses

Geographical Scale	**1991 Census**	**1991 Census**	**2001 Census**	**2001 Census**
	Areas	Data set name	Areas	Data set name
Postcode	1991 postcode units	ED/PC postcode directory	2001 postcode units	Headcounts for postcode units
Small Area	Enumeration Districts (E, W, N); Output Areas (S)	Small Area Statistics	Output Areas (UK)	Census Area Statistics
Ward	1991 electoral wards & equivalent (E, W, N); 1991 postcode sectors (S)	Local Base Statistics	2001 electoral wards or 2002/3 electoral wards; varies by data set in small ways	Standard Tables
Local Authority	1991 Local authorities	Local Base Statistics	2001 Local Authorities	Standard Tables

- **British Household Panel Survey (BHPS) (ESDS, ULSC)** The BHPS started with the first wave in 1991 and the latest available wave is Wave 13 (2003). A sample of 5,500 households and 10,300 individuals was collected for the first wave from 250 areas of Great Britain. The BHPS added more coverage to Scotland and Wales at wave 9 in 1999 by adding 1500 more households to each of the two countries, which only had around 400-500 households each before that. An additional sample,

2,000 household records, was collected from Northern Ireland, which formed the Northern Ireland Household Panel Survey (NIHPS), so that the panel became suitable for UK-wide research at Wave 11 in 2001. The BHPS allows social and economic changes to be closely monitored, interviewing the adult members in the households annually. Analysis is possible for Great Britain from waves 1 to 10 and for the United Kingdom from Wave 11 onwards. The BHPS covers more aspects and more details of households and the lives of household members than the Census.

3.2.2. Simulation Processes

Processes to be simulated are designed according to the objectives of the model (O'Donoghue 2001). The objectives should be well presented so that essential processes can be structured from them (Zaidi and Rake 2001). The processes are commonly designed and written as modules. This approach has the advantage of great flexibility and extensibility, and is probably influenced by modern programming techniques such as object-oriented languages. In principle, a dynamic model includes a few demographic modules in order to 'evolve' the start population into the future. The demographic modules include at least ageing, mortality and fertility. Many models have more modules: coupling to create new families and decoupling to demolish existing family units, young people leaving home to create an independent family from their parents, emigration and immigration. Most microsimulation models are proposed and developed as socio-economic models. Therefore, education and employment may be not only treated as significant social processes in their own right but also be the input components for modelling economic components such as income. A spatial microsimulation model requires at least an inter-region migration module and in more advanced versions an intra-region migration module. Various types of methods have been invented and explored for modelling these processes based on micro units.

First of all, the time frameworks used in microsimulation models are of two types: discrete time intervals and continuous time. Discrete time models refer to those models updating population annually. Time is always discrete in a computer. Therefore, the term 'continuous' in principle refers to a time unit smaller than year, which is a month in the DYNAMOD model (Galler 1997; King *et al.* 1999; Bækgaard 2002). The

problem of competing risks, caused by multiple mutually exclusive events simulated in the same time, is possible in discrete model (Galler 1997; van Imhoff and Post 1998). For example, whether death of a female or giving birth by the female occurs first will be a competing risk for the possible infant since both of them can be possibly simulated in the same year though only one will happen. That can be avoided by continuous time models by executing the closest event within the year, in which if death of the mother happens first the birth will then be aborted. Continuous time models can have probabilistic approaches to update variables monthly. They more commonly relate to applying survival functions. That starts the second feature of microsimulation models introduced in the next paragraph.

Applying survival functions will generate a list of events with their occurrence time for each micro unit. At the time an event happens, the involved micro unit transits to a new state so the model can also be called an 'event driven' model. Besides this, the second feature includes two more methods for modelling processes: behavioural and probabilistic approaches. The three dynamic methods are briefly described here:

- **Behavioural modelling** originated from economists' attempts to update the behaviour of micro units in reaction to the change of institutional characteristics such as a tax policy which might result in reducing labour supply. This kind of dynamic has been classified as second order effects by Bekkering (1995). The first order effects are static effects, which is the objective of a static model, such as the income tax recalculation caused by the changes of tax rates while other variables remain constant. The third order, cyclical effects are illustrated by a dynamic balance of supply and demand (Bekkering 1995). A labour supply that is rising may be caused by a tax policy and then a salary decrease might result in turn. After that demand for labour may be stimulated so that supply will drop. Bekkering (1995) also suggested that it is practically difficult to apply the third order effects in microsimulation models whereas it is easy to analyze these at the meso and macro level. O'Donoghue (2001) noticed that very few dynamic models have designed even the second order effects.

- **Probabilistic modelling** is the most common technique applied to dynamic population into the future, involving simulating mortality and fertility with the

assistance of Monte Carlo simulation. For example, if the probability of dying for a 80-year old male is 0.5 and if a random number (in the range 0 to 1) less than 0.5 is given by executing a random number algorithm, then the person will die, whereas he will be still alive if the random number is more than or equal to 0.5. The probabilities can be achieved in two ways. Transition matrices are the easiest to use though it may be difficult to derive them. For example, mortality rates by age, sex, marital status can be achieved from single published life tables. However, if mortality probabilities further disaggregated by social class and employment are needed, various estimation techniques may be required. In order to incorporate more variables in determining the transition processes, regression may be more preferable. Logistic regressions are commonly applied to derive the transition probabilities from empirical data. For instance, death records can be pulled out from available longitudinal data on which logit regression can be run to produce coefficients for selected determinants such as age, sex, marital status, social class, job and health condition.

- The best example to explain the use of **survival and hazard functions** in microsimulation is mortality. Mortality rates are typical hazard rates $h(t_h)$, which are defined as the probability that a component (person in this case) will fail (die) just after time t_h given that this person has survived up to time t_h. That is obvious for mortality since a person cannot die more than once. If time *t* is a continuous variable, then a survival function $S(t)$ is defined from hazard rates as

$$S(t) = \exp(-\int_0^t h(t_h)t_h) \tag{3.1}$$

(for the way this is derived, please see pp296-300 of Pitman (1993), where t_h is the continuous time element in the hazard function. The discrete time form of this equation is

$$S(t) = \exp(-\sum_{i=1}^{n} h_i t_i) = \prod_{i=1}^{n} S_i \tag{3.2}$$

where the time *t* is discretely divided as *n* intervals t_i and hazard rate h_i is assumed constant within the time interval t_i. It can be noticed that the discrete time equation will be equal to the one for continuous time when the t_i closes to *0*. In other words, the smaller time intervals are used, the better approximation the discrete time equation will

achieve compared to the continuous time form. It may be at least part of the reason that continuous time microsimulation models use monthly time units instead of years. In the following, time *t* is assumed as a discrete time unit. Two common methods are applied to generate events to drive a microsimulation model in the continuous time framework. In the first method, survival functions $S(t \mid X)$ can be directly achieved from published life tables for a specific event such as death, where *t* is the time and X is the vector of one or more explanatory characteristics such as sex and social class for mortality. For example, at the time of birth, the death date for this new person can be calculated by finding out t from $u = S(t \mid X)$, where u is a random number between (0, 1). The characteristics of this person will determine a survival function to be used, and then the random number *u* will be generated as the survival rate for this person. Finally the corresponding time *t* could be found in the *u* survival rate position. Another way is to derive survival functions using hazard models estimated by hazard regression from empirical data. The common hazard function used is a piecewise exponential hazard model, which assumes hazard functions follow an exponential distribution

$$h(t) = h_b \exp(\beta(t)X(t)) \tag{3.3}$$

in which h_b is the baseline hazard rate, $X(t)$ is the explanatory variables at time *t* and $\beta(t)$ is the coefficients for $X(t)$, both of them are assumed constant within a time interval for the ease to handle this model and together they are the covariates determining the hazard rate. In principle, it is a simplified form of Cox's proportional hazard model

$$h(t) = h_b(t) \exp(\beta(t)X(t)) \tag{3.4}$$

in which the baseline hazard $h_b(t)$ is variant rather than invariant as in equation (3.3). Again, the reason for that is easier to handle this model.

$$S(t \mid X) = \exp(-\sum_{i=1}^{n} t_i h_b \exp(\beta(t_i)X(t_i)) \tag{3.5}$$

can then be derived by replacing the h_i in equation (3.3) with equation (3.2), then another form

$$\ln S(t \mid X) = -\sum_{i=1}^{n} t_i h_b \exp(\beta(t_i)X(t_i) \tag{3.6}$$

can be achieved by putting a $\log_e$ on both sides. Then in the similar way as the first method does, time *t* can be calculated by assigning a random number *u* as the survival rate $S(t \mid X)$ in (3.5) or (3.6):

$$\ln u = -\sum_{i=1}^{n} t_i h_b \exp(\beta(t_i) X(t_i) \quad (3.7).$$

Klevmarken (1997) argues that all of the three modelling methods can be classified as behavioural modelling. Further, he also suggested that data imputation could be treated as another kind of behaviour, which imputes missing variable X_2 from another data source having both X_2 and X_1 by matching X_1 in both data.

The third feature that distinguishes microsimulation models is the nature of system closure. Models can be designed to be open or closed. The term 'closed' is rather confusing, since researchers may mistreat it, constructing a closed model which will not have processes of bearing children and importing new migrants. Closed actually only refers to the selection of partners from existing individuals in a coupling process rather than the creation of a new individual. An open model has two major benefits. When the modelled population size is small, an appropriate partner may be difficult to find among the existing individuals. If modelling in a closed manner, unrealistic couples will be generated, for example, a couple with a huge age difference. Therefore, an open model is preferable to create partners with realistic characteristics. The second merit is that a open model can be executed in parallel on a machine with multiple processors or a computer cluster since individuals can be simulated without interacting with other individuals. The trade off is increased complexity and difficulty in alignment which may consume the time saved from running the model in parallel, and as a result most dynamic models in use are closed (O'Donoghue 2001).

The fourth feature of microsimulation models is the degree to which input variables are allowed to vary. Models can run with fixed system parameters such as stable tax and economic growth rates into the future, which is called steady state, compared to models forecasting the future population with possible dynamic trends of these parameters. The main advantage for utilizing a steady state model is that it allows the effects of the current characteristics of a system to be isolated and investigated without being complicated and confused by various effects caused by varying behaviour.

These techniques are commonly applied in combination rather alone, for example, an open steady state cohort model using transition probabilities or a closed population projection model utilising the survival functions.

3.2.3. Alignment

Bækgaard (2002) presents a thorough investigation in which he recoganised that the objective of alignment is to compensate for imperfectness of data and estimation techniques. However, he also reckons that the right way to correct the result of a wrongly designed process is to redesign or adjust the process rather than alignment. Nevertheless, it is still recommended that alignment is able to input additional information for the compensation of the output resulting from insufficient data estimation.

- Not only the total output but also the distribution of base data and output can be aligned. Further, alignment not only includes the calibration of processes but also the adjustment of base data caused by, for example, sample stratification. The input data of DYNACAN (Morrison and Dussault 2000) have been adjusted in many ways, one of which is the reweighting of some elderly samples and also changing the ages and genders of selected elderly samples to fit a known aggregation. The Combinatorial Optimisation approach may be understood as an alignment to adjust national samples to fit known small area aggregates.

- The alignment to the processes will be specified by what kind of microsimulation method is involved such as transition matrices, regression, survival functions or hazard functions. Nevertheless, two classes of alignment can be recognised: alignment of output and alignment of parameters.

- For the probabilistic approach, both transition matrices and regression functions, the alignment of an output can be implemented by Monte Carlo sampling more micro units if the output is less than the constraint, and abandoning those sampled micro units if they are more than required. A more advanced approach is to rank sampled micro units in the order of their probabilities from either transition

matrices or regression functions and then transitions will be applied to these units from higher probabilities until sufficient number of transitions applied.

- The alignment of parameters for transition matrices can be a process that trial Monte Carlo sampling is firstly simulated and the ratio of the simulated to the target number can then be used to adjust the probabilities. For regression functions, it is the parameters in the function to be adjusted by the ratio, which could be a complicated process if it is not a simple linear function.

- Three methods have been tried for the alignment of survival functions. The first is adding a time-based adjustment factor which is computed by comparing the simulated aggregated number of events with the target number of a specific year, to the baseline hazard value. This way was proved to be laborious and complicated since the adjustment will be tried at the start of the simulation and the whole process including the simulation needs to be restarted again and again until the achievement of the right values. The second method is adjusting the baseline hazard value in a similar way as the first method does but period by period in the simulation rather than at the start of the simulation. In the third method, the survival function will be used to produce a list of events which will be ranked by their dates from the start to the end of the simulation. Then in every period, the target number of events will be selected from the list to be executed in this period without concerning the planned dates from applying survival function.

Further the alignment process should also be applied in execution of scenarios based on different assumptions about the future, which can be implemented by the alignment of the processes using the assumptions.

The alignment process should link household microsimulation models to other parts of the 'world system'. More ambitious than using predefined values as constraints, an untested way is to link microsimulation models to a larger 'world system' model. Commonly, a microsimulation model only models household sectors. Then the output of household microsimulation model will be the input to other parts of the 'world system' model and then the larger model will react to the input by sending signals to the microsimulation model as new constraints.

The alignment process should reduce Monte Carlo variance. The aggregation of a process employing Monte Carlo sampling may vary significantly especially when the sample size is small. A common way is to run the model many times and aggregate on the average outcome. An alternative way is using alignment, which has been discussed above.

3.2.4. Policy Analysis and Scenarios

Demographic scenarios commonly specify different demographic time series rates to determine the future population such as the gradual decrease of fertility and mortality rates. They could be implemented in microsimulation models in another way. For example, the decrease of fertility rate is caused partly by the increase of female labour participation. Therefore, if the labour market is implemented in a microsimulation model, the labouring process will influence fertility indirectly.

Microsimulation was originated to analyse the impact of policy changes. The implementation varies according to the field in which the model is applied such as tax-benefit reform. Policy impacts could be included in scenarios, since scenarios are also designed to answer questions such as "what will happen if assumption 1 or assumption 2 is adopted". They are implemented in the simulation of behavioural processes (O'Donoghue 2001), which has been described in section 3.2.2. The SESIM (Pylkkänen 2001) model which models wages and hours of work is an example of a second order of behaviour, based on which possible policy scenarios can be implemented. The parameters of utility functions to compute the reacting behaviour for different household types are estimated from LINDA (Longitudinal Individual Data for Sweden). The labour supply of individuals can be adjusted in reaction to economic environment changes such as income tax change. The effects on hours of work from an income tax change can be divided into an income effect and a substitution effect, which generate two opposing effects on the labour supply. If an income tax change results an increase in hourly wage, the income effect increases the demand for leisure based on the assumption that leisure has an income elasticity of more than one, which reduces the labour hours to the market, whereas the substitution effect stimulates the individuals to work more when they want to compensate for the change in income due to the tax change. The total effect is decided by the combination of the two effects. Two policy

changes, doubling the personal tax exemptions and using a flat tax rate instead of progressive part of the taxation, are experimented with to study the behavioural response from single mother and two-adult households (Pylkkänen, 2001).

3.3. Spatial Microdata Reconstruction

One of the first tasks in the construction of a dynamic microsimulation model is to select a suitable micro dataset that can represent the attributes of the base population, and the suitability of a dataset used as a base can be best judged in relation to the objectives set out for the model (Zaidi and Scott 2001), which holds true for the static models as well. A fairly complex and time consuming process is involved to generate the 'ready to use' micro data in many models (for example, Duley 1989; Williamson *et al.* 1998), since full population micro data are available in only a few countries such as Sweden (Holm *et al.* 2003).

Richer and richer micro data are being produced in the United Kingdom from both government surveys, such as the Census, General Household Survey, Labour Force Survey, and the private sector. For example, the company NDL (National Demographics and Life style) (Williamson *et al.* 1998), has detailed demographic and 'lifestyle' information for over half of the country's households. Information about population presented in the micro data form has the merits of efficient representation, flexible aggregation, and data linkage. Unfortunately, the available micro data is classified at a relatively coarse level of spatial resolution in order to protect the confidentiality of the respondents. Nevertheless, another form of population data, the census tabulations of population, were produced in more spatial detail, down to the Enumeration District (ED) level at 1991 and Output (OA) level for 2001. The Census provides the most authoritative social accounting of people and housing in Britain and is a unique source of data for the social sciences (Dale and Marsh 1993).

A number of different approaches to the creation of spatially detailed micro data have been developed, including data fusion, synthetic reconstruction (chain imputation) and reweighting. Data fusion and merging requires levels of access to the original micro data not normally permissible due to the protection of respondent confidentiality, whilst shortcomings in published small-area data mean that conventional stratified sampling is

unable to capture the highly complex and multi-dimensional nature of between-area differences (Voas and Willamson 2001). Synthetic reconstruction has been used in UK when suitable micro data have not been available (Birkin and Clarke 1988; Duley 1989; Williamson 1992; Birkin and Clarke 1995). Combinatorial Optimisation has also been applied in various projects such as SimLeeds (Ballas *et al.* 2005c), David Simmonds Consultancy (Ballas *et al.* 2005b) and by Williamson *et al.* (2002).

3.3.1. Synthetic Reconstruction

The synthetic reconstruction approach is the most long-standing method for generating synthetic microdata. It normally involves Monte Carlo sampling from a full joint probability of the characteristics of households and individuals to create imaginary lists (Williamson *et al.* 1998) of individuals and households. The full joint probability commonly does not exist but separated conditional probabilities are possibly derived from published tabulations, such as Small Area Statistics (1991 Census) or Standard Area Statistics (2001 Census), and iterative proportional fitting (IPF) is deployed to estimate possible links among them.

The process commonly starts with generating the right number of household heads and their base attributes, for example, age, sex, marital status in an area from a known distribution, then with the assistance of IPF, the distribution of economic activity could be added to these heads. If they are economically active, then in a similar way, occupation and industry could be assigned. If married, their spouse as well children will be generated. In a sequential way, the attributes required will be attached to them step by step.

As an example, the population reconstruction work in Duley's doctoral thesis (Duley 1989) begins by generating age, sex, marital status for the head of household constrained by 1981 Census SAS Table 26, and further disaggregated at age and marital status by using 1981 Census District statistics Table 35. The combined widowed/divorced category is broken up into separate widowed and divorced categories using 1981 Census Table 7. District data disaggregate the age into more categories, and then national data is used to deconsolidate them into single years. In order to attach ethnic and country of birth (COB) to the heads, three marginal probabilities, which are

derived from COB national Country of Birth Volume Table 1 by marital status by sex, COB by age by sex national Country of Birth Volume Table 2 and age by marital status by sex 1981 Census Table 6, are merged using IPF since the full joint probability of age, sex, marital status and ethnic does not exist. In addition, as ethnic status has to be based on the COB of the head (in the 1981 Census though not in later censuses) their place of birth characteristics can be generated by definition. In a similar fashion, family status, living arrangement, economic activity, socio-economic group, employment status and industrial group are generated.

Williamson *et al.* (1998) suggested that previously unknown relationships between various attributes could be derived from SARs and used to generate new probabilities which would strengthen weak links in the reconstruction process. Therefore, in Huang and Williamson (2001), the conditional joint distribution estimations are derived through a three-level estimation procedure using IPF in order to generate their synthetic population.

- The national level joint distribution for the all variables is derived directly from the SAR.

- The ward-level joint distribution will be estimated from ward level constraints and the national level joint distribution as the initial estimates using IPF.

- Similarly, the ward-level joint distribution will serve as the initial estimates in the IPF estimation process for the ED-level joint distribution.

The process could be summarised as iteratively adjusting the elements of an array of a higher geographical level to fit known constraints of the lower geographical level guided by IPF, such as from ward to ED, from nation to ward etc., so that the interaction pattern of the higher level one will be retained at lower levels. They (Huang and Williamson 2001) also concluded that the decisions on what and how many other attributes should be included in our synthetic data and the ordering of their generation are guided by the perceived importance of a variable in 'determining' others (Birkin and Clarke, 1998), the availability of suitable local data linking a variable with those already created, data quality and cost.

In the Huang and Williamson application, three main predictors of headship, age, sex and marital status are selected from the logistic regression analysis of SAR which analyses the significances of various factors to predict headship as the first step. Then, household head with age, sex and marital status can be generated from 1991 Census SAS table35, then with 1991 Census SAS table39 non-heads by age, sex and marital status can be generated as well. In addition, the ages of the heads are disaggregated into finer categories. Then, the rest of the attributes are synthetically reconstructed in an order determined primarily by data availability.

The major problems of synthetic reconstruction approach are as follows.

- The sampling error from deploying Monte Carlo sampling for small areas is likely to be significant since the sample sizes are small, which are about 200 households or 450 people at ED level. The geographical scale in Duley's doctoral thesis (Duley 1989) is postcode sectors and he adopts a threshold of 1000 population. If a postcode sector has a population less than the threshold, the whole postal district containing the postcode sector will be treated as one small area.

- Synthetic reconstruction is a sequential procedure so that error introduced in each stage by various reasons will accumulate through the chain of generation of characteristics. The order of the sequential procedure may help to minimise the error, but, because of the lack of an appropriate 'scientific' approach the determination of the ordering relies on the modeller's skills and art (Birkin and Clarke, 1995).

- Subjective choices still need to be made about the relationships between variables where no obvious data are available at the small area level. Data at higher geography levels need to be involved, though the detail of data at small area level is getting better.

3.3.2. Combinatorial Optimisation

The combinatorial optimisation approach is an attempt to select a combination of households from the microdata source, such as the Sample of Anonymised Records (SAR) to reproduce the characteristics of a chosen area as far as possible, in which a

household might be selected or cloned more than once. If one possible combination of SAR households is a solution in the whole solution space that represents all possible combinations, then a computer algorithm is required to seek the optimal solution since the solution space defined by the quantity of the micro data and the dimensions of the characteristics is in general too huge to be searched by a simple exhaustive searching method. Iteratively swapping of one or more households within the combination guided by the algorithm is the general process to achieve the optimal solution. Williamson *et al*. (1998) implemented and compared a few algorithms which concluded that Simulated Annealing is the best algorithm to use though it is computationally intensive.

Solutions were evaluated in terms of the discrepancies between the summary or aggregation tables of a solution and published population statistics such as the 1991 Census SAS, which are employed as an abstract of Census capturing the area characteristics. The summary tables are constructed in the exact same format as the constraint tables for the convenience of the evaluations. The simplest evaluative criterion is a measure of the total absolute differences between the summary and the constraint tables, which are the errors. Voas and Willamson (Williamson *et al.* 1998; Voas and Willamson 2001) have reviewed a dozen statistical measures and concluded that the most suitable for assessing the fit of synthetic microdata to published small-area constraints was the normal Z-score and related variants.

In principle only a small subset of the SAS is used as the target or constraint tables mainly due to the constraints of computer resources and the amount time for model development. Guidance has been provided in Voas and Willamson (2000). First of all, the interesting variables for the topic should be included as far as possible unless they are not in the SAS, since although the constraints chosen will determine the values of unconstrained variable by which they are correlated to each other, the extent of this determination will never be better than moderate (Williamson *et al.* 1998). Secondly, attention is required to include as many variables as possible in as few tables as possible. It not only reduces CPU time but also provides the necessary and critical linkages between variables. What synthetic reconstruction and combinatorial optimisation do is actually the same thing, that is, estimating a full joint probability out of many possible ways of joining published separate conditional probabilities, but in different ways. Thereafter, if a published table has provided a joint probability of three variables,

estimating another joint probability of the three variables from two separate probabilities, each of which is for two of the three variables, is actually the wrong way, since the estimated result is only one of the many possible linkages which may not be identical to the true and available joint probability.

When involving more than one constraint table, as reported in Williamson *et al*. (1998) and found in other cases, errors are concentrated in one or some tables. For example, the output of a university area does not fit well for the term time address table. The poorly fitting tables are typically those where the actual statistics reflect a distribution very different from the national norm, and hence from the SAR (Voas and Willamson 2001). The reason might be that the other normal distribution tables will converge much more rapidly than the term time address table. Once a near-optimal combination for the normal distribution tables is reached, it becomes very difficult to accept any changes. The solution proposed is to fit the abnormal tables first, but then to accept any subsequent overall improvement only if it does not degrade the position in relation to that individual target (Voas and Willamson 2001). Huang and Williamson (2001) suggested that:

- Total absolute error is used to evaluate the solutions during Simulated Annealing and the evaluation of the fit to the known small area tables is based on Z score. However, an alternative statistic would improve estimates and the extra cost still needs to be explored.

- Using regional subsets from the SAR for the majority of EDs where characteristics are close to the norm is quite feasible since the households in each subset may better reflect regional differences in household characteristics not constrained in the household selection process, which has been done by other projects such as SimLeeds (Ballas *et al.* 2005c) and David Simmonds consultancy project (Ballas *et al.* 2005b). However, for atypical EDs only using the regional SAR would significantly increase the error of estimation, and so that the whole SAR should be used.

- If using sequential table fitting approach, the order of tables to be fitted is area-specific, and the target level of acceptance is table-specific. Then applying this

approach will be difficult in generating large area micro data, since so many small areas are required to be processed in their own orders of tables.

3.4. Review of Selected Microsimulation Models

Reviews of microsimulation models have been presented by many authors. For example, Merz (1991) focuses on static models and a survey of dynamic microsimulation modellings can be found in O'Donoghue (2001). Further, Van Imhoff and Post (1998) summarised the merits of microsimulation models in a thorough comparison of microsimulation models of demographic processes with equivalent macrosimulation models. This section intends to provide details of a number of selected dynamic and static microsimulation models as examples of the principles discussed in the section 3.2.

3.4.1. SVERIGE Spatial Microsimulation Model

The review is based on Holm *et al.* (2003) and Rephann *et al.* (2005). SVERIGE is the first national interregional spatial dynamic microsimulation model based on households with individual members, which starts from a unique large socio-economic database, TOPSWING. This database contains longitudinal information about every person living in Sweden between 1985 and 1995 covering the topics of demography, work, family, income, employment, transfers and location. Spatially, the micro units are georeferenced to squares of 100 by 100 metres and then the squares are located in the LARegions system which divides Sweden into 108 different labour market regions. The labour market region is treated in the same way as domestic addresses. The model is built in a closed dynamic discrete time framework which runs in single year intervals and contains modules simulating the processes of ageing, mortality, fertility, emigration, education, marriage, leaving home, divorce, migration, immigration, and employment and earnings.

The model is programmed in C++ to update the characteristics of persons in every simulation year. Monte Carlo simulation is implemented to determine the occurrence of specific events in a person's life by sampling from estimated transition matrices or probabilities calculated from logistic equations derived from TOPSWING. Some of the changes in characteristics are triggered by predefined rules such as changing the marital

status from married to widowed when a partner dies. The use of transition matrices may be illustrated by the immigration module. The immigrants are selected from an immigration pool having 60,000 individuals. The demographic distribution of heads is determined by a constant transition matrix and another transition matrix is used to assign the immigrants to a labour market. Both matrices are derived from the more than 100,000 immigrant records. Four equations are applied to determine the probability of dying. For example, for persons from 25 to 59 year old, the probabilities are calculated by a linear equation with parameters derived from regression analyses using age, sex, education and other variables as predictors.

Intra-regional or inter-regional migrants are modelled separately. Intra-regional migrants are only modelled by a number of simple rules for the life events of cohabiting, divorce and leaving home. Three steps are involved to migrate households to other regions which are: decision to move, choice of region and allocation of 100 metre square. Stepwise logistic regression is used to estimate the probability of the decision to move, which is influenced by a set of variables such as the age, education, earning and employment of the head, age of oldest and youngest child, how long the head of the family has stayed in the same dwelling, number of previous moves and immigration group. The choice of a target labour market region is based on a multinomial logit regression with separate parameter estimates for each origin. The probabilities for each possible move from a specific labour market are calculated for every labour market based on regional variables such as distance, average earnings, and vacancies in labour market. Allocation of a migrant household to a 100 metre square is decided by the closest matching of the migrant to possible destination squares according to a compatibility index based on earning, education level and family size. However, the lack of a housing module reduces the accuracy of intra-region movements because the model fails to include the availability of opportunities at the destination.

The model runs a module that reduces family members first through mortality and emigration, and then a fertility module that adds family members. After that the rest of modules are executed in the order of education, marriage, leaving home, divorce, migration, employment, earnings and immigration. Experiments showed that model outputs had not been significantly affected by some reordering of the modules (Rephann *et al.* 2005).

The simulation from 1990 to 2000 without alignment results in a substantial underestimation of total population, which may be caused by the underestimated observed fertility during the first half of the 1990s. Further, part of the underestimation of population growth is due to overestimates of mortality. Immigration is designed as policy instrument: the total immigration figure is input as absolute number.

With macro alignment, the total outputs of those modules are close to the input from alignments, hence the investigation of performance focuses the relative distribution among regions, which is entirely model determined. Observed and simulated distribution of total population match quite well. However, the regions with the smallest populations are slightly underestimated and most regions with larger populations are overestimated. The observed population is in general well matched with the simulated population by sex and age, and population by age and education level, with part of the distributions mismatching. The incomes are underestimated. However, this can be corrected by applying an adjusting factor. The immigrations are underestimated for the smaller regions and overestimated for the larger regions. The births and deaths are in general well simulated with small overestimation of births for the regions containing low population. A relatively high overestimation of emigration is simulated for the regions which are sparsely populated. The migrants are overestimated from and to most regions.

3.4.2. UPDATE: Spatial Dynamic Microsimulation Model for Updating Census

UPDATE was developed in School of Geography, University of Leeds as a PhD project to update Census information during the 10 year interval at a small area scale in the United Kingdom (Duley 1989). The geography selected for small area modelling is postcode sectors for commercial marketing purpose. A method to map postcode geography to Enumeration Districts (ED) had been developed by Pinpoint Analysis Ltd using grid referencing of the centroids of postcode units and the coordinates of ED boundaries.

The base spatial micro population is generated by the synthetic reconstruction approach using 1991 small area statistics data, which has been introduced in section 3.3. Multi-regional migration was abandoned becaused of the difficulty of handling a huge small area migration matrix. A simplified bi-regional approach is applied instead in which wholly moving households and individual migrants shift their residences between a postcode sector and an external region consisting of the rest of the world. The relocations of eligible migrants are driven by housing provision and requirement. Mortality, fertility, pair formation and pair dissolution are modelled using local transition probability matrices. It is noticeable that significant proportion of the thesis and probably large proportion of the modelling time have been dedicated to estimations or even guestimations of the data without the benefit of the excellent data available to Swedish social scientists mentioned in the last model review. Only births and deaths statistics are published annually down to ward level. IPF was extensively employed to estimate data for ward level from statistics for larger geographic areas. Surveys such as the annual General Household Survey and National Health Survey Central Register were used to compensate the shortage of high coverage panel data for cohabiting and migration respectively. The variables involved in each module were limited by data availability so that the heterogeneities within the subgroups of population can not be explored. For example, deaths are modelled based on age and sex so that the differential mortality behaviour among social classes were not captured; the influences of education and labour participation on fertility were not incorporated.

The model has been applied for updating only six postal sectors for validation, because of the limits to the power of computing hardware in the late 1980s. The use of data from higher geography caused the deviation of the projection from the target since the data did not capture local variations.

3.4.3. Baldini: Dynamic Cohort Microsimulation Model (Baldini 2001)

Dynamic cohort or longitudinal microsimulation models belong to a sub-type of dynamic microsimulation models, in which a cohort of micro units born in the same period are followed from the birth to the death of the last member. This type of model is

mainly applied in projects requiring a life-cycle analysis, which is longer than the usual few decades of dynamic simulation. In a study of inequality and redistribution, Baldini (2001) pointed out that the level of current income for micro units may greatly differ from that of accumulated income during their whole life; current income is influenced by temporal life shocking events which can give a wrong picture of usual living standards. The increasing social mobility in an open society also requires long term study of the income positions, though existing panel data may not be sufficiently long.

Therefore, Baldini (2001) starts a dynamic cohort microsimulation with a total synthetic population of 2000 men and 2000 women representing the Italian population though he also suggested that the population of the cohort could also be derived from a 'real' micro data pool. The dynamic process updates the population for 95 years using constant probabilities to present its demographic and economic features assuming the characteristics of Italy do not change, so that the characteristics of the current welfare system could be isolated and investigated. The probabilities of the demographic and economic events are derived from official publications and surveys. All demographic modules, mortality, education, marriage, divorce and fertility, are processed by Monte Carlo sampling from the derived probabilities. The economic modules, labour supply, hours of work and earning, use transition matrices, regression functions or equations based on rules.

The demographic modules for a cohort model do not aim at producing a representative future history of the population. Overall, the simulation proved that incomes should be less unequally distributed in a life-cycle setting than an annual framework.

3.4.4. SAGE: Simulating Social Policy in an Ageing Society (Cheesbrough and Scott 2003)

SAGE is a closed microsimulation model using household SARs as base data and models all of the events on an annual discrete-time framework. The demographic part does not implement internal migration since the model is not spatial and immigration or emigration is not included in the prototype of the model. The demographic modules are

executed in the order of mortality, fertility, partnership dissolution, cohabitation formation and then marriage.

The probabilities of mortality by age, sex and social class are collected by using single year probabilities from vital registration by single year of age and sex in 1991 to disaggregate the grouped probabilities from the abridged life table provided by the ONS Longitudinal Study.

The probabilities for fertility and partnership transition modules are both derived from BHPS data. Logistic regression analysis of BHPS data from 1992 to 1996 produced two fertility predictor equations for women living with a partner and without a partner respectively. The equation for women with a partner is based on the woman's partnership stratus and duration of cohabitation, the age of the woman, the number of children and the age of youngest. Only the latter two factors are included in the equation for women without a partner.

Partnership dissolution and formation are all modelled as female dominant and the processes for cohabiting and married couples are simulated separately. The logistic regression models for the possibility of cohabitation dissolving include woman's age at formation of cohabitation, duration of cohabitation, whether there is a birth in the next year, whether the woman previously was married and whether either partner is a full time student. For divorce, the equation is basically the same except that the factor whether man is 6 or more years younger than his partner and their duration of cohabitation is less then or equal to 9 years, replaces the factor whether either partner is a full time student.

Logistic regression analysis of the BHPS is used to estimate the probabilities of the transitions to cohabitation and marriage based on a group of variables such age, marital status, education status and the duration of cohabitation. Eligible women enter the partnership market to find a partner. The partnership market is processed in two steps:

- A conditional logistic regression model is used to analyse the characteristics of a combination of true new couples from 700 ONS General Household Survey new couples records and 24 false couples each of which is the woman of the new true couple with a randomly selected fake partner.

- For a woman entering the partnership market, the probabilities of coupling can be calculated for any possible partners by inputting their combined characteristics in the conditional logistic regression. Then a Monte Carlo process is utilised to select one of them as the true partner.

A simple education module is adapted in the prototype, which only distinguishes two education statuses: being in full-time education and having higher education qualifications. People aged 16 to 21 years are simulated to stay in education until they leave, among which students will be qualified if they can stay until age 21. The probabilities of leaving by age were estimated from the participation rates of each age in the SAR.

3.4.5. SimBritain: A Spatial Microsimulation of Full Britain Population (Ballas et al. 2005a)

SimBritain was developed at University of Leeds initially and at later stage the research team moved to University of Sheffield. The full population of Britain was constructed and modelled in SimBritain at the parliamentary constituency level using data from British Household Panel Survey (BHPS) and 1991 Census small area statistics (SAS). BHPS data were chosen over SAR data even though the BHPS is smaller, since the modellers think that the longitudinal nature of BHPS data is essential to study the social and economic change through time. A pilot model, SimYork, was developed to test different methodologies and combinations of datasets, which micro modelled the population of York city at ward level.

SimBritain is a static spatial micro model which projects the 1991 small area statistics into 2001, 2011 and 2021 employing Holt's linear exponential smoothing (Holt 2004) to extend the trend from the 1971, 1981 and 1991 Censuses small area statistics into the future. SimBritain then adjusts the weights of BHPS micro data to fit the small area statistics and their projections guided by Iterative Proportional Fitting. The weights are adjusted iteratively by the ratio of the cell values in the aggregation tables of the synthetic population to the corresponding cell values in the corresponding SAS tables

until the aggregation converges to the SAS tables. Solutions were experimented to tackle the problem of the York wards populated with non-Northern households in the SimYork project. However, none of them works well hence, so it is suggested by the modeller that it is better to only use the micro data from the same areas or nearby areas.

The linear exponential smoothing is validated by comparing the projected 1991 distribution from 1961-1971-1981 Censuses with the actual 1991 distribution. The modellers suggest that the project method works relatively well for most areas except a few local authorities.

3.4.6. DYNAMOD: A Continuous Time Microsimulation Model (Galler 1997; King *et al.* 1999; Kelly and King 2001; Bækgaard 2002)

DYNAMOD is a sophisticated continuous time microsimulation model in which survival and hazard functions are applied for some modules to generate the date of events updating population into the future. It was developed at the National Centre for Social and Economic Modelling (NATSEM) in Australia. It is interesting to notice that the development of the first version of DYNAMOD was interrupted by the resignation of the entire DYNAMOD team, because of the complication of employing too many untested methodologies such as an internal macroeconomic model and survival functions. As a result, the internal macroeconomic model was abandoned and the survival functions were only applied in some of the modules including fertility, mortality, disability, couple formation and dissolution. Probabilistic approaches using monthly intervals are employed in some of the modules.

The input base population is the 1% sample from the 1986 Australian Census, which has 150,000 individual records organised in family units. Census variables are not sufficient for the modelling purpose so that data imputation is applied to bring in more characteristics, including state of residence, disability status, level of current education course, earnings, and a number of variables about people, family, education and labour force histories. Data other than the base data includes fertility, disability and mortality

rates, emigrant parameters, labour force targets, earnings parameters and new immigrants.

Mortality is modelled by applying survival functions derived from actuarial tables based on age, sex, disability status and year of birth. Dates of death are calculated from the survival functions at birth or the start of the simulation, and will be recalculated when the disability status changes since it is the only variable that is able to change. The disability is also modelled by applying survival functions estimated from 1993 survey about disability.

The birth dates are determined for females for different types of birth such as pre-marital birth, first marital birth and second or subsequent marital birth using the corresponding survival functions at the beginning of the simulation and in January of each year. The survival functions are defined by piecewise exponential hazard models. Thereafter, fertility is modelled in an annual schema which allows alignment with published birth data. The survival functions were derived from 1986 survey data based on the age of woman, her fertility history, educational participation and attainment, labour force status and marital status, together with the labour force status of her partner where applicable.

Cloned existing immigrants and constructed new cases are input into the pools of potential immigrants. The weights of these records are adjusted according to the specifications of the required distributions of particular characteristics. Cases are then selected every year though they are brought into the model evenly over the 12 months. Emigrants are selected with reference to target numbers and the age and sex distributions of emigrants aged 18 years or over, and then removed from the model in the similar way of the introduction of immigration into the model.

Transition probabilities are applied annually in January to model young person leaving home. The transition probabilities are generated from logistic regression models based a group of variables including age, sex, state of residence, type of school attended is used to analyse and predict whether a person leaves home.

Couple formation and dissolution are modelled using 10 survival functions for different types of formation or dissolution such first cohabitation before marriage and first marriage, which are estimated in the same way and from the same data source as those for modelling fertility. The survival functions are also defined by piecewise exponential hazard models. For example, the date on which a female starts the first cohabitation is calculated using a survival function based on her current type of educational institution and whether she is pregnant, whereas the separation date is calculated using a regression equation. Cohabitation, marriage, and the both types of couple dissolution are female dominant. Possible mismatches are avoided in the case of couple formation if they are modelled separately, while in the case of couple dissolution the survival functions have incorporated the characteristics of both partners. A male partner is randomly selected with reference to the age and educational qualifications of both partners, and to the labour force status of the male.

Modules using annual transition probabilities are aligned by two methods. The first method is to adjust a transition probability by the ratio of target number of transition to the trial simulated number using the pre-adjusted transitions probability. The adjusted transitions probability is then be applied to do the 'real' transitions. The second method will generate an individual list in the order of the transition probability values generated for the individuals in the list. Then individuals will be selected from the list according to the target number and the probabilities from high to low.

The methods to apply survival functions and align their output are the major innovations this model achieved. In the third version DYNAMOD, three ways are used to align survival functions, which have been introduced in section 3.2.3.

The alignments not only introduce a more plausible model but also provide the capability for scenario setting. The overall performance of the aligned model is good in which the actual total Australian population for the period 1986-2000 and the most probable projection of the 2001-2050 population matches the simulated population well, and the projected population by age structure is good as well.

3.4.7. DYNACAN: a Canadian Discrete Time Dynamic Microsimulation Model (Morrison and Dussault 2000)

DYNACAN was built to analyse Canadian Social Security Schemes such as the Canada Pension Plan (CPP). It was designed primarily to project the trend of the CPP using its current configurations, and also to allow assessment of the impacts of policy changes to the CPP. DYANCAN is a typical discrete time dynamic microsimulation model in a sense that it models all demographic and economic modules by applying Monte Carlo simulation using probabilities, which are derived from administrative, Census and survey data, user assumptions, and the assumption of the CPP Actuarial Valuation Model. In addition, the projections of DYNACAN are required to be consistent with the output of the CPP's Actuarial Valuation Model, a macrosimulation model.

DYNACAN starts with an input micro data base which contains 212,935 individuals organized in families and generated by adjusting micro data published from 1971 Census. The demographic part of the model has emigration, immigration, a bi-regional internal migration between Quebec and the Rest of Canada, mortality, fertility, aging, marriage, divorce and leaving home. A typical trial and adjustment alignment was applied to these modules using the probabilistic transitions. More specifically, the probabilities are adjusted by adjustment factors calculated from the difference between the target number and the number of events generated in the trial run.

3.4.8. Hägerstrand Migration Model

Hägerstrand (1957) developed a multi agent model in order to explore the relationship between social contact and migration. It is a classic and brilliant example of spatial demographic simulation since it was done by hand before computing power could be utilised by researchers. It may be worth noting that the focus of reviewing this model is not the migration but the spatial demographic simulation modelling principles.

The model is founded on two key definitions. Firstly, a migration field is devised, in which population and vacancies are evenly distributed, and the area is divided into square cells of equal size. Secondly, migrants are divided into active and passive types.

The active migrants can choose a destination in a randomly selected adjacent cell, whereas the passive migrants are stimulated by the active migrants in a way that a passive migrant chooses a new cell where an earlier migrant from the same origin is staying and the attractiveness of all earlier migrants are equal. The second definition means migrants follow the path of earlier migrants which is the principle Hägerstrand intended to examine.

For processing the simulation model successfully by hand, a few simplifications were introduced to this model in order to limit the processing work which include the two key definitions. First of all, population and vacancies do not distribute evenly in a real world so that the attractiveness of each earlier migrant is different, and active migrants can only choose one of the surrounding cells to move. Secondly, only one origin cell can generate new migrants which are only 50 individuals every year and they will not return to origin unless other rules lead to it. The third, passive migrants will be passive only once, after that, they will be active migrants for the rest of time, which means only the social contacts established in the origin are important and counted. What is more, migrants having left the origin for more than 25 years, can no longer attract passive migrants from the origin anymore. Further, the model disregards migrant mortality.

Nevertheless, it still has some variations. Firstly, the more times a migrant moved, the less the possibility of the migrant will move again. Secondly, the probability of movement decreases when they stay in a cell longer. Finally, the probabilities of the migrants from origin being active and passive are 40% and 60% respectively.

The process could be simply divided into two phases. First of all, the first and active migrant moves to one of the nearby cells randomly selected. Then, the type of next migrant will be decided by the associated rules. If it is passive, then where a random selected earlier migrant stays will be the cell to move; otherwise it will migrate according to the active way.

The model produced fairly sophisticated pattern of migrants although it has the limitation and only run for 25 years since computers were not commonly available to researchers at that time. Hägerstrand also introduced a few modifications to improve the model. For example, evenly distributed vacancies could be evolved to unevenly

distributed vacancies with vacancies of variable attractiveness. Another important field to explore is that the variation of the proportion of active and passive migrants and the spatial range of the active migrations.

Nevertheless, the principles of building a demographic simulation model could be derived from this example, though a real case will be much more complicated than that. First of all, the basic entities, time interval and the geography should be defined. In this case the basic entity is an individual migrant, geography is square cells of equal size and the time interval is a year. Secondly, 'fuel' is required to run this model which could be defined as the processes built into the model, which change the demography of this region. One of the processes in this example involves a social contact about the vacancies in those cells, so that an earlier migrant could stimulate a passive migrant from the origin to fill a vacancy. That implies the third principle that is the rules to link causes and the change of demography, which could be seen as how to burn the 'fuel' to run the 'engine'. In this case it includes the rules to change the attributes of the migrants according to those probabilities, possible variation of demography which could be the variation of vacancies. A common method to decide the attributes of the entities according to the attribute distribution is Monte Carlo Sampling.

3.5. Outline of a Prototype Spatial Dynamic Microsimulation Model for the David Simmonds Consultancy (Ballas *et al.*, 2005b)

A prototype of spatial dynamic microsimulation model has been developed in School of Geography for David Simmonds Consultancy, which hosts a household location and transportation project commissioned by the UK Department for Transport. This microsimulation model is part of the project to replace the macro household location component in their DELTA model, which models residential location and transport. The forecasts of the microsimulation will be used as inputs to transport modelling.

This prototype is a spatial, dynamic, discrete time and transition matrices based microsimulation model, which is implemented in Java, modelling demographic and housing events in the full population of Bradford and Craven, Calderdale and Kirklees,

Leeds and Wakefield, Barnsley, Rotherham, Sheffield, and North East Derbyshire, Doncaster and Bassetlaw, for 283 wards from 1991 to 2001 and from 2001 onwards. The idea of running the model from 1991 to 2001 is to test and validate the model using existing historical data from 1991 to 2001. In addition, various software tools have been built to process and analyse the data and outputs. The model can be divided into three parts: data processing, demographic modelling and housing modules.

3.5.1. Input Data Processing

Two sets of base micro data were generated for 1991 and 2001 using Combinatorial Optimisation implemented in a Java program, that has elements of a generic population reconstruction system, which will be discussed in Chapter 4. The program is coded using message passing interface (MPI) so that it can fully utilise the power of a Beowulf Cluster which has 54 CPUs in total. The Combinatorial Optimisation takes 1991 Household SARs from the Yorkshire and Humberside, and East Midlands regions, and 10 SAS tables as constraints to generate micro population for the 1991 microsimulation and 7 SAS constraint tables for 2001. The constraint tables from the 1991 Census are:

- SAS 42: Household composition and housing
- SAS 35: Age, sex and marital status of household residents
- SAS 8: Economic position
- SAS 57: Household space type, rooms and household size
- SAS 9: Sex, economic position and ethnic group
- SAS 73: Industry
- SAS 74: Occupation
- SAS 10: Term-time address
- SAS 84: Highest qualification, age and economic position
- SAS 21: Car Ownership

The reason that micro data construction for 2001 has only 7 tables is that 3 of the proposed 10 tables were not available in CASWEB when the model code was written and 1991 Household SARs had to be used for 2001 because the 2001 Household SARs had not yet published. The constraint tables from the 2001 Census used are as follows:

- CAS 062: Household composition by number of cars or vans available
- CAS 002: Age by sex and marital status

- CAS 032: Sex and age and level of qualifications by economic activity
- CAS 050: Dwelling type and accommodation type by tenure
- CAS 036: Sex and industry by age
- CAS 039: Occupation by industry
- CAS 061: Tenure and car or van availability by economic activity
- CAS 064: Households with full-time students away from home and age of students by number of students, households with students away from home
- CAS 113: Occupation by highest level of qualification
- UV 09: Ethnic group

Both 1991 and 2001 areal statistic census data were extracted using the CASWEB interface supported by the Census Dissemination Unit (CDU) at the University of Manchester.

The micro data generated for 1991 fits the constraints fairly well. The discrepancies between the aggregation of the micro data and the constraints are not much for most the wards though wards around university areas have high errors. For 2001, the fit of the microsimulated micro data to the 2001 Census tables is far lower than its fit to the 1991 Census tables. The first factor suspected is that the characteristics of the population in 2001 are rather different from those of 1991 population; therefore 1991 micro data samples may not represent the current population well. Further, some 2001 Census variables have different categories or definitions from those of 1991 Census such as industry and ethnicity. A compromise was reached to aggregate some 2001 Census categories and map them as rationally as possible to their counterpart in 1991 Census categories. After that, additional variables were imputed to the micro data sets such as work placement from Census journey to work matrices and income from BHPS.

3.5.2. Demographic Microsimulation

Ageing, mortality, fertility, couple formation and separation are included in the dynamic demographic component. Except ageing, all the modules are processed using transition probability matrices. However, these probabilities are not location specified, which means local variations have not been incorporated. The reason may be that the client requires the flexibility to run the model in other regions without the pain of replacing

location-specific parameters. The mortality probabilities are disaggregated by age and sex, and then weighted by social class. The fertility rates are disaggregated by age and marital status, and then weighted by ethnicity. Decoupling only models divorce, which is a male dominant process using divorce rates by age. The dissolutions of cohabiting couples are ignored. Marriages are divided into two kinds: singles getting married and marriage by cohabiting. Singles getting married is also a male dominant process. Eligible males and females are collected and then single male marriage rate are applied to the eligible males. After that, a male processed by the Monte Carlo simulation will be matched with a random selected eligible female which is decided by a set of deterministic rules involving the difference of their ages. The cohabiting couple to married couple transition is executed using male dominant probabilities by age. However, the number and distribution of cohabiting couples are not constrained in the Combinatorial Optimisation process since SAS does not have a table for that, although the SAR has variables indicating cohabiting. The cohabiting couples reduced by marriage will be augmented by the single to cohabiting process using the similar procedure and rules as in the single to marriage process. It is assumed that cohabiting numbers will be stable over years.

3.5.3. Household Location

Part of the household relocation has been implemented in the coupling and decoupling since new household need to be formed in both modules. A set of rules will decide who is going to move out in decoupling or how to merge two partners. The initial idea to use social class to drive relocation in principle was abandoned since the variations of migrants across social classes are not significant from the analysis of BHPS. Instead 'housing stress' was explored in which a *person per room ratio* is calculated. The analysis of BHPS based on 'housing stress' shows a higher *person per room ratio* gives a higher probability to migrate and also the average room change by the ratio has been derived. Then eligible migrant households are collected using the probabilities by occupancy and then the destinations are decided by rules about potential destination's deprivation score and the change of household income.

Other researchers have developed the rest of the model, which includes modules of enter labour market/stay in education, leave/re-enter/retire from labour market, update

educational status, update qualification, fix/change occupation, driving licence, absence from household, household location, employment *etc.*

3.6. Lessons Learned and Design for the Construction of a Microsimulation Model for Water Demand

As a result of the literature review and the building of the prototype microsimulation model for the David Simmonds Consultancy project, lessons have been learned for both model structure design and technical implementation, which will be discussed in section 3.6.1 and 3.6.2 respectively. The design of the water demand microsimulation model will be introduced based on these lessons and the objectives of this project in section 3.6.3. Basically microsimulation brings great advantages over macro modelling; other WaND team members had experimented with macro approaches so micro modelling approaches were a complementary choice. According to the discussion in Table 3.1, housing-led population projection for small area scale was a preferable option for this project, modelling residential water demand, for which accommodation type is a key variable. Further, housing is the main driver for migration at Middle level Super Output Area (MSOA) scale. Finally, choosing static rather dynamic microsimulation approach is trade off caused by the limitation of research time. However the projection of households provided by Parsons *et al.* (2007) will enable the modelling of water demand at micro household level for selected future years.

3.6.1. Lessons for Model Structure Design

The following remarks apply mainly to dynamic microsimulation models but many are also relevant when constructing static microsimulation models

- Clear objectives and the associated essential processes need to be identified at the start of modelling and other modules can be added in a latter stage (Zaidi and Rake 2001). Further, being too ambitious about the sophistication of modules will result in failures. Demographic modules are essential to most models, but education and labour market are not so frequently included. Simplified education and labour market modules can be implemented for presenting a complete model and they can be ‘refurnished’ if time and cost allow it.

- In a spatial context, the size and scale of the study area is critical and needs to be carefully decided. Ballas *et al.* (2005d) constructed a pilot model, SimYork, to experiment with the model design and implementation at region level first. Then SimBritain (2005a) was constructed on a sparser geography since it is for the whole of Britain.

- An advanced migration module might be desirable, but a simplified migrant process using the existing migrant data will be preferable at the initial stage.

- Using a macro model to interact directly with a microsimulation model might be too ambitious to be successfully implemented in a normal research period, 3 years or less. Therefore, it is preferable to input the projection output from the macro model for UK local authority population and household projections into micro model in an alignment mechanism.

- A traditional discrete time framework might be better for the ease of implementation, which will model processes in an annual schema. Using monthly time units can be experimented within a later stage. Data availability restricts the use of survival functions and monthly time units so that it is more reasonable to model some processes using conventional transition probabilities on an annual time unit (King *et al.* 1999)

- UK does not have a database as valuable as LINDA in Sweden, therefore the model should reflect the limitations of UK data. Although Census provides the LS, it contains only a 1% sample and has 10-year intervals.

- Alignment and the validation using historical data will be essential. Historical data need to be collected both for total population and vital statistics. Not only the distribution of population but also the outputs of modules need to be validated.

- Should transition matrices or regression functions or survival functions be used? The choice might be determined by the availabilities of the data and the objectives

of the model. In principle, transition matrices are preferable where published transition probabilities are available, whereas regression is not a bad choice when those probabilities need to be estimated from empirical data. Further, regression is more convenient when the determinants of the probabilities are many. Applying survival functions is conceptually elegant and can achieve more efficient computation. For example, the date of death event can be only calculated once through the whole microsimulation process. However, this statement is debateable. First of all, deriving the survival functions is much more complicated than preparing probabilities, which certainly costs substantial modelling time. Secondly, the calculation of event dates such as death could occur more often than that in theory. In DYNAMOD, the mortality is only disaggregated by age, sex and disability; therefore obviously not many times of calculation are required since only disability will change and it seldom changes. However, if the mortality is determined by age, sex, social class and marital status or more variables, then the number of recalculations may be raised by social mobility and the changes of marital status. Furthermore, the alignment for applying survival functions are much more complicated and requires substantial computation time as well, therefore it may cancel out the amount of time saved by doing so. None of these factors is as serious as the requirement that a spatial microsimulation model incorporates local variations. Therefore, either probabilistic regression or survival functions need to be localised for so many small areas, which has not yet be done by any of the existing models.

In conclusion, probability matrices may be preferable where the module requires localised parameters. Regression might be suitable for incorporating many variables and few large regions rather than many small areas. Survival functions are preferable at the stage when the model fairly complete applying other methods. Then the replacement of some of the probabilistic processes can be explored if time and costs permit.

- Synthetic Reconstruction or Combinatorial Optimisation. A full and thorough assessment and comparison of the two approaches has been implemented (Huang and Williamson 2001) and concluded that:
    - Combinatorial optimisation has the great advantage of the flexibility in selecting constraint tables and the ability to generate individuals within

families and households at the same time, whereas synthetic reconstruction needs to generate data in a predefined order and generate members of households separately.

- The synthetic microdata produced by the methods, using the same small area constraints, fit constraining tables very well. But the microdata from combinatorial optimisation is less variable than the output of synthetic reconstruction, at both ED and ward levels.
- In terms of programming, synthetic reconstruction is also more complicated and time-consuming. In other words, the ease of combinatorial optimisation modelling might be another important gain for some modellers. Although combinatorial optimisation is computing intensive, Moore's law is making this feature a less and less of a problem.

In addition, it is argued (Ballas *et al.* 1999) that another merit of combinatorial optimisation over synthetic reconstruction is the former approach uses the 'real' people. In the light of above, combinatorial optimisation is preferred because of its flexibility, better quality of output microdata and the ease of modelling.

- The construction of the input base data with all variables required involves more than one step and method, since a data source seldom contains all the necessary variables. Thereafter, other variables can be attached to the microdata generated from synthetic reconstruction or combinatorial optimisation by data imputation, which can be performed by regression functions, closest distance matching or a more intelligent way such as a neural network.

3.6.2. Lessons for Implementation

- It is convenient to debug and test the code with a small proportion of the whole base dataset since reading a few million records from a ASCII file is time consuming and the same process probably need to be executed hundreds if not thousands of times.

- A flexible program may be necessary to read parameters from an input file rather than hard coded them in the code. It will be easy to interpret the model and change parameters for many purposes for example scenarios.

- If the raw base data does not have all the variables required, they may be imputed from other data source by matching the deterministic variables between them.

- Code should be well documented as well as the model design.

- Programming language. The nature of microsimulation is in favour of object oriented language and also it requires speed since processing of millions record is computing intensive. Researchers commonly list a few options, suggest many criteria for the language of microsimulation, and have put forward many arguments for and against each language. However, eventually the language they chose was the one they were skilled at even though it did not fit all the criteria. The language chosen for this study is Java since it is object-oriented. However, Java is not as fast as C++ and consumes more memory. Nevertheless, Java is cross-platform, which means programs written in Java can be developed in a user friendly platform and executed in a UNIX or LINUX machine with more power without changes. In addition, the author is skilled in Java so that modelling time will be significantly reduced.

- A few years ago, memory was so insufficient that microsimulation modellers became experts at caching data, which stores most micro units on disk rather than all in memory and only reads in the records required in the simulation process. Having 2G bytes memories in a 32 bit PC, several million records can be read at once and contained in the memory which not only reduces the processing time but also eases the development process. However, the limit does not seem far away. If a model is required to cover whole country or have tens variables if not hundreds, then a caching mechanism is still required. Database software might be a good alternative, which has to be fast to query and update millions records.

- Only one person or one family or one household is required in processing an event in most modules so that processing one unit at a time can be designed if memory runs out. The module requires more than one units is commonly the coupling and decoupling, for which a special routine could be designed.

3.6.3. Model Design

The microsimulation model is proposed to contain the following modules and implementation:

- Reconstructing micro residential population in Thames Gateway 2001 using 2001 SAR and CAS by implementing Combinatorial Optimisation and Simulated Annealing algorithm;
- Aligning the micro population from 2001 to 2031 in yearly interval for matching with the population projection from Parsons *et al.* (2007), which becomes a static projection module of the microsimulation.
- Specifying micro units, households and individuals with common demographic attributes but also water demand relevant attributes;
- Statistical matching the resident population records from a with Domestic Consumption Monitor so that the micro population can be attached with water using related variables such as number of water using appliances, metering and water consumption volume.
- Exploring to the use of the macro household projection model developed by Parsons *et al.* (2007) as the input for the static microsimulation projection of this thesis modelling, which is a household population matrix by household size and accommodation type in Local Authority resolution. The microsimulation model will be dynamic its household population in the same categories to match the projection output from the macro model. The projection of household population is a mix of expert opinions and target figures from authority planning reports which includes demolition and refurbishing of the existing houses and, construction of new houses.
- Executing scenarios about new developments including new housing, new technologies for water conservation, and new behaviour by consumers facing incentives and disincentives.

3.7. Summary

Population modelling has been discussed in the dimensions of macro or micro and housing-led or population-led. The chapter concludes that a microsimulation model commonly involves assembling a base of micro data, a simulation process to make

dynamic the micro units and alignment to validate the output. Through these steps, scenarios can be implemented and policy impacts can be evaluated. A spatial static microsimulation approach to combine the power of spatial microsimulation and macrosimulation has been proposed based on reviews of related techniques, applications and development of a prototype microsimulation. Details of microsimulation modelling are examined in the review of selected applications. Then lessons learnt were presented and a clear modelling structure has been set for the implementation of this modelling work.

The next chapter will discuss key techniques and presents a package of software implementing these techniques for the modelling.