Creating a Synthetic Population

Rolf MOECKEL
Research Associate
Institute of Spatial Planning
University of Dortmund
August-Schmidt-Strasse 6
D-44221 Dortmund
Germany
Tel: +49-231-755 2127
Fax: +49-231-755 4788
E-mail: rm@irpud.rp.uni-dortmund.de

Klaus SPIEKERMANN
Partner
Spiekermann & Wegener
Urban and Regional Research (S&W)
Lindemannstrasse 10
D-44137 Dortmund
Germany
Tel: +49-231-1899439
Fax: +49-231-1899443
E-mail: ks@spiekermann-wegener.de

Michael WEGENER
Professor
Spiekermann & Wegener
Urban and Regional Research (S&W)
Lindemannstrasse 10
D-44137 Dortmund
Germany
Tel: +49-231-1899441
Fax: +49-231-1899443
E-mail: mw@spiekermann-wegener.de

Abstract

There is a new interest in integrated models of urban land use and transport provoked by the environmental debate. However, most existing urban models are too aggregate to respond to these challenges. New activity-based models require more detailed information on household demographics and employment characteristics. New neighbourhood-scale planning policies to promote the use of public transport, walking and cycling require more detailed information on the precise location of activities. The method for this new type of model is Monte Carlo microsimulation. Theses models aim at reproducing human behaviour at the individual level, i.e. how individuals choose between options following their perceptions, preferences and habits subject to constraints, such as uncertainty, lack of information, and limits in time and money.

Microsimulation models require micro data. However, the collection of individual micro data, i.e. data that can be associated with single buildings, or the retrieval of individual micro data from administrative registers is neither allowed in most countries nor desirable for privacy reasons. Therefore these models work with synthetic micro data that can be retrieved from general accessible aggregate data. A synthetic population has to be generated that represents individual actors in the form of households and household members. A synthetic population is statistically equivalent to a real population. For each household characteristics such as household size, income, number of cars and address are generated. Each person is described by characteristics such as age, sex, religion, and work location. For creating addresses for the synthetic population, land-use data are disaggregated to raster data by GIS techniques.

This paper gives a theoretical introduction and presents two applications to generating a synthetic population. While most existing procedures concentrate on iterative proportional fitting the approach to generate synthetic populations presented here combines different approaches. The study area for the first application was the city of Netanya in Israel. The main process to generate a synthetic population is shown. The second application was the metropolitan area of Dortmund with a population of about 2.6 million. Businesses, dwellings, and non-residential floorspace are generated, too.

Keywords: Synthetic Population, Microsimulation, Micro Data

1. INTRODUCTION

There is a new interest in integrated models of urban land use and transport provoked by the environmental debate. However, most existing urban models are too aggregate to respond to these challenges. New activity-based models require more detailed information on household demographics and employment characteristics. New neighbourhood-scale planning policies to promote the use of public transport, walking and cycling require more detailed information on the precise location of activities.

Also travel models are confronted with new challenges to urban modelling. New alternatives like park-and-ride, car-sharing, or shared buses and new life styles or new work patterns cannot be simulated with traditional four-step travel models. Aggregate travel models are unable to reproduce the complex spatial behaviour of individuals and to respond to sophisticated travel demand management measures. Furthermore the attention paid to environmental aspects of transport requires more detailed information about emissions caused by different scenarios. As a reaction new activity-based travel models have become more popular, although they require more detailed information on household demographics and employment characteristics. As the availability of more powerful computers has overcome former barriers to handle large data bases, new disaggregated travel models aim at a one-to-one reproduction of the behaviour of individuals (Salomon et al. 1998, Wegener 1999).

The method for this new type of model is Monte Carlo microsimulation. Microsimulation models aim at reproducing human behaviour at the individual level, i.e. how individuals choose between options following their perceptions, preferences and habits subject to constraints, such as uncertainty, lack of information, and limits in time and money. Microsimulation models require micro data. However, the collection of individual micro data, i.e. data that can be associated with single buildings, or the retrieval of individual micro data from administrative registers is neither allowed in most countries nor desirable for privacy reasons. Therefore these models work with synthetic micro data that can be retrieved from generally accessible aggregate data.

The synthetic population represents individual actors of the model in the form of households and household members. Each household has certain characteristics like household size, income, number of cars and address. Household locations determine some of the travel origins and destinations. In addition, each person is simulated by characteristics such as age, sex, nationality, and employment. Persons are assigned activities they accomplish, they go to work, go to school, shop or go to the doctor. They choose a transport mode and so produce traffic flows. In the long run, households may decide to move and so affect land use. The synthetic businesses represent the employers in the model. Businesses are described by their industry, number of employees, number of vehicles and location within the study region. Public amenities are a special case of businesses. They include institutions like kindergartens, schools, universities or museums. They affect land use by the foundation, relocation or shutdown of businesses. The synthetic population further consists of dwellings where households are living, and non-residential floorspace where businesses are located. In other words, actors and the built environment are generated in order to build a synthetic urban setting for running microsimulation models. Since this population only exists artificially in files on the computer, it is called synthetic. It is statistically equivalent to the real population of the study area.

This paper first gives a theoretical introduction to microsimulation and micro data (Sections 2 and 3). After that a synthetic population developed for the city of Netanya in Israel is presented to explain the main procedure to generate a synthetic population (Section 4). Finally an introduction to an ongoing application for the Dortmund region in Germany is given to show the level of detail a synthetic population can obtain and its future development (Sections 5 and 6).

2. STATE OF THE ART

2.1 Microsimulation in Spatial Models

Microsimulation models aim at reproducing human behaviour at the individual level, i.e. how individuals choose between options following their perceptions, preferences and habits subject to constraints, such as uncertainty, lack of information and limits in disposable time and money. Basically microsimulation is the reproduction of a macro process by many micro processes. Single events of distinct actors are the basic building block of microsimulation. No deterministic assertions that are valid with certainty can be made. Instead probabilistic assertions that are valid only with probability are made about events.

Microsimulation was first used in social-science applications by Orcutt et al. (1961), yet applications in a spatial context remained occasional experiments without deeper impact, though covering a wide range of phenomena such as spatial diffusion (Hägerstrand 1968), urban development (Chapin 1974, Chapin and Weiss 1968), transport behaviour (Kreibich 1979), demographic and household dynamics (Clarke et al. 1980, Clarke 1981, Clarke and Holm 1987, Holm et al. 2000) and housing choice (Kain and Apgar 1985, Wegener 1985, Mackett 1990a, 1990b). In recent years microsimulation has found new interest because of its flexibility to model processes that cannot be modelled in the aggregate (Clarke 1996). Today there are several microsimulation models of urban land use and transport under development in North America: the California Urban Futures (CUF) Model at the University of California at Berkeley (Landis and Zhang 1998a, 1998b), the Integrated Land Use, Transport and Environment (ILUTE) model at Canadian universities under the leadership of the University of Toronto (Miller 2001), the Urban Simulation (UrbanSim) model at the University of Washington, Seattle (Waddell 2000), and the 'second-generation' model of the Transport and Land-Use Model Integration Program (TLUMIP) of the Department of Transportation of the State of Oregon, USA (Hunt et al. 2001). There are no efforts of comparable size in Europe. There are a few national projects, such as the Learning-Based Transportation Oriented Simulations System (ALBATROSS) of Dutch universities (Arentze and Timmermans 2000) or the Integrated Land-Use Modelling and Transportation System Simulation (ILUMASS) in Germany. The generation of the synthetic population for ILUMASS is described in Section 5.

2.2 Synthetic Populations

The most prominent approach of generating a synthetic population was developed by Beckman et al. (1996). Today this procedure is used in most integrated models that generate a synthetic population. Basically, iterative proportional fitting was used to estimate the proportion of households in a zone with a desired combination of demographic aspects. Since corresponding input data were available, three different household types were generated separately: family households, non-family households and group quarters such as college dormitories or prisons. Microcensus data available for the United States were used as input (Public Use Microdata Samples, or PUMS). Local aggregate data were used as row and column totals. The number of households for each matrix cell was obtained by multiplying the total number of households by the probabilities in the estimated table or by drawing a number at random according to these probabilities. In the first case special attention had to be paid to correcting round-off errors. The microdata had 11,760 cells of which 11,151 were zero entries. These cells might not be empty in the real population, therefore empty cells were set to 0.1 or 0.01 before using the iterative proportional fitting routine.

This procedure to generate a synthetic population was first applied in the TRANSIMS project (Los Alamos National Laboratory 2003) which aims at creating a new integrated transport and air quality forecasting procedure. The system permits the modeller to modify the behaviour of each individual. Since travel is influenced by age, income, gender and employment

status both persons and households are generated. Land-use data are used to place each household on the transport network. Finally cars by emission type are assigned to households. In addition fictitious persons are generated that represent those who travel on the network but live outside of the study area.

A variant of the procedure developed by Beckman et al. was used to generate a synthetic population for UrbanSim, a model simulating urban development for land-use, transport and environmental planning (Waddell 2002). Each household is represented by household size, age of head of household, presence of children, income and number of workers. Employment is represented as individual records for each job and its employment sector. Jobs are located in non-residential floorspace or in residential buildings to account for home-based employment. Households are linked with individual dwellings and individual jobs.

A rather similar approach was used for generating a synthetic population in the land-use transport model of the State of Oregon in the United States (Hunt et al. 2001). Households are taken through a demographic transition including ageing of the household members, birth, deaths, departure of members to form new households and possible dissolution of the household. The transition probabilities are functions of various location and socio-economic factors and are adjusted so as to respect the marginal distribution for the total population. The household income and qualifications to work are updated, too. If a new household is formed using members departing from existing households the household and person characteristics for these households are drawn by randomly selecting values from known marginal distributions.

Ton and Hensher proposed a procedure to generate a synthetic population of 401 households representing the 1.5 million households of Sydney (Ton and Hensher 2001). Each household is described by socio-economic and demographic characteristics and a weight representing its contribution to the total population. For instance, if a household carries a weight of 200, this indicates that this household represented 200 actual households in the total population. Through time the base year weights are modified to represent the changing composition of households in the population.

The synthetic population developed by Hertkorn and Wagner (2002) consists of 1 million persons living in 509,000 households in Cologne, Germany. By iterative proportional fitting persons by 16 different person types are generated. These persons are associated with households depending on the household size. A node of the transport network was sampled as housing location.

Several other projects to generate synthetic populations are in progress. The Integrated Land Use, Transportation, Environment (ILUTE) modelling system aims at developing a synthetic population for a large urban microsimulation land-use transport model (Miller 1996). The ALBATROSS microsimulation model, a 'learning-based' transport oriented simulation system, generates a synthetic population for a base year and under scenario conditions (Arentze and Timmermans 2000). There are further applications that represent actors as agents (Torrens 2001, Veldhuisen et al. 2000, Schelhorn et al. 1999, Batty et al. 1998). Rule-based agents do not have individual characteristics but all members of a group are equal. Agents are used for cellular automata models.

3. METHOD

For generating synthetic populations aggregate socio-economic data are disaggregated by bi-proportional and multi-proportional adjustment. By additional information like digital land-use plots or aerial photographs synthetic input data are produced and located which are

statistically equivalent with the aggregate data. Three methods are crucial for generating a synthetic population: iterative proportional fitting to transform one-dimensional into multi-dimensional data, Monte-Carlo sampling and disaggregation of zonal data to raster data.

3.1 Iterative Proportional Fitting

In general, statistical input data are available only as one-dimensional distributions for zones such as census tracts or statistical areas. This could be populations by age, households by size or businesses by business sector. In order to generate a synthetic population, multidimensional distributions are required, such as persons by age, gender and education or households by size, nationality and income. One-dimensional distributions are transformed to multi-dimensional distributions by iterative proportional fitting (Deming and Stephan 1940). This approach is identical to the RAS method used in input-output analysis to adjust a matrix of intersectoral flows to known regional inputs and outputs. It alters the elements of a matrix in such a way that the sum of rows and the sum of columns equals the known one-dimensional input data, and that the deviation from the initial values of the matrix is minimal.

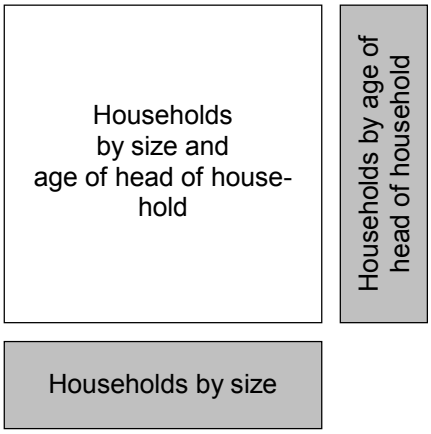

Figure 1 Iterative proportional fitting

For instance, to generate a synthetic population, information about the relationship of household size and age of the head of household is required. Since younger persons tend to be head of smaller households while 45-year old persons tend to be the head of larger family households, it is crucial to know the joint probability of household size and age of the head of household. Data provided by the administrative registers usually are one-dimensional, i.e. the distribution of size of households and the distribution of age of head of household are given separately. These distributions form the sum of rows and columns in Figure 1. If available, initial values of the fields of the matrix can be filled with data from older censuses or from comparable regions. If no appropriate data are available, the initial values are guesstimated. Iterative proportional fitting estimates the number of households by size and age of head of household for every field of the matrix so that the column and row totals equal the one-dimensional input data.

The method of iterative proportional fitting can be applied to two- or multi-dimensional matrices. For instance, when generating the synthetic population of dwellings in the Dortmund region (see Section 5) a four-dimensional matrix, containing information about dwelling size, dwelling quality, ownership, and kind of building was set up. For each of the four features one-dimensional data were available from the administrative registers. As initial values for the four-dimensional matrix, data from the 1968 housing census were used. This procedure was used to generate the housing stock for the year 2000.

3.2 Monte-Carlo Sampling

When generating more complex sets of micro data the iterative proportional fitting has its limits. It implies very high standards in terms of reliability of initial input data for the cells of the matrix. Detailed micro input data like PUMS in the US are not available everywhere. If features have to be generated that are not covered by the micro data other procedures have to be found to generate those. The iterative proportional fitting has to set zero-cells to 0.1 or 0.01. This influences probabilities and is theoretically hardly founded.

Monte-Carlo sampling allows generating an almost infinite number of different features. With Monte-Carlo sampling synthetic multidimensional data are gained from one-dimensional distributions of administrative registers (Wilson and Pownall 1976). Figure 2 visualises this procedure. To generate a household, first the age of head of household is selected. Depending on that information the household size is selected. A younger head of household tends to live in a smaller household, while a 45-year old head of household tends to live in a larger family household. In other words, households and persons are generated in a 'natural' order. This means that the features of persons and households are sampled in the order they influence each other. Because every step of the sampling refers to the already selected features, the result becomes more realistic. In the example of Figure 2 the likelihood for selecting the household size are adjusted according to the age of the head of household. The probabilities are determined by iterative proportional fitting. After all household members have been determined, the number of cars is selected. Again, the probability for the number of cars depends on the already selected features, i.e. on the age of head of household, the household size and the age of further household members. This process implies a causal structure for the order of the generated features.

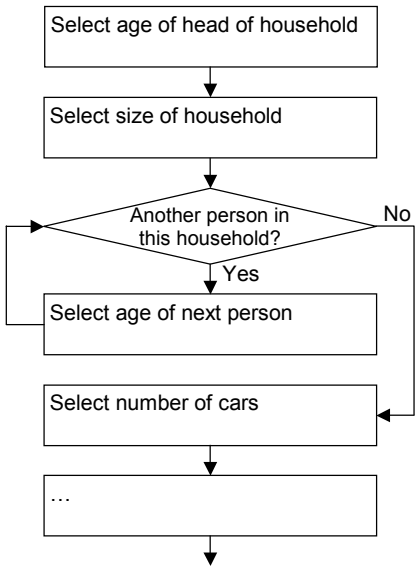

Figure 2 Monte-Carlo Sampling

With Monte-Carlo sampling as many features can be selected as necessary for running a microsimulation model. The number of features is only limited by the possibility of determining reasonable relationships between the selected features.

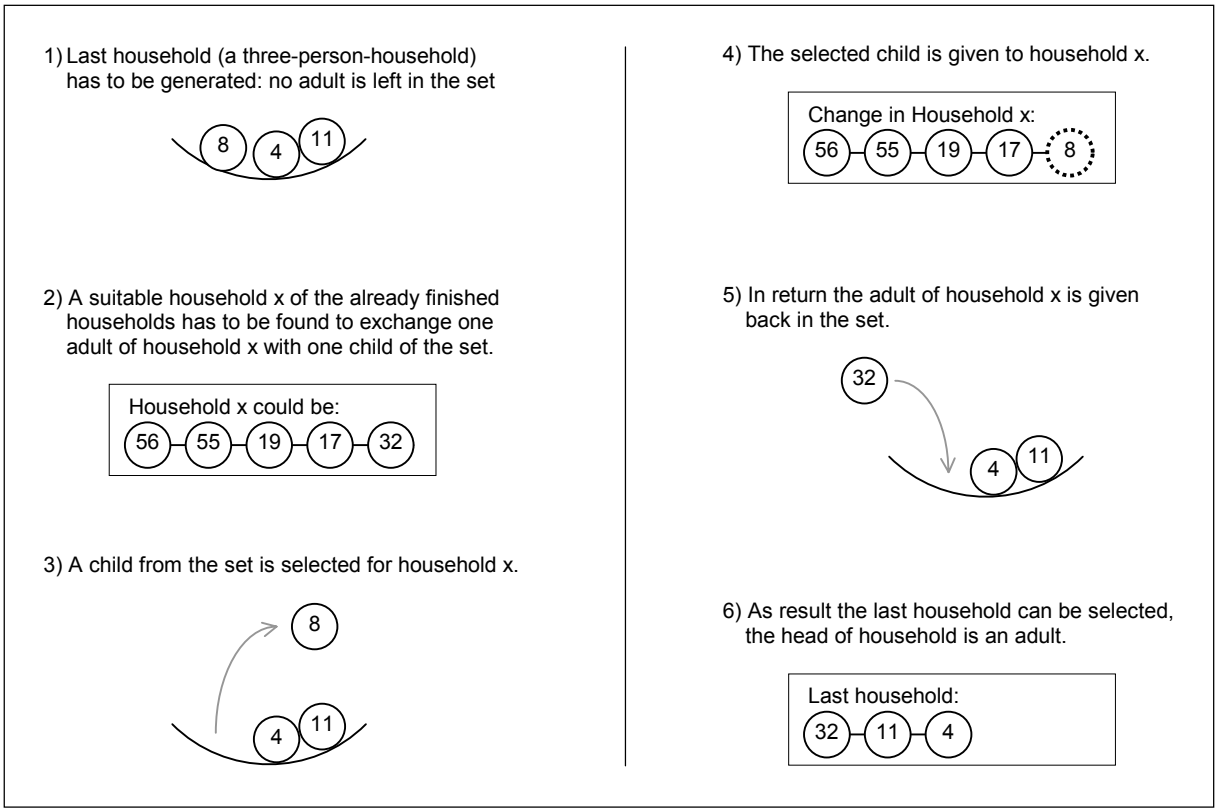

Figure 3 Procedure of a submodel to exchange persons

During sampling most selections are 'drawn without replacement', i.e. features like number of persons by age and gender, number of households by size or number of cars remain exactly as given by the administrative registers. If these features were sampled by 'drawing with replacement', there would be only approximately 5.2 percent of the population boys under 5 years of age. 'Drawing without replacement' assures that there are exactly 52 boys under 5 in 1,000 persons, i.e. it eliminates a random error.

A difficulty is that at the end of the simulation there are only few persons to be selected left. For instance it might happen that at the end of the simulation of a zone a three-person household is to be generated but there are only three children aged 11, 8 and 4 years left (Figure 3). These persons cannot form a reasonable household. An adult is missing to finish this household. In this case a special subprogram is run to exchange one adult from an already finished household and exchange it by one of the children left in the set. In this way the last household can be generated with one adult and two children.

3.3 Disaggregation of Spatial Data

Activity-based microsimulation transportation models require the exact spatial location of activities, i.e. geographic co-ordinates or point addresses as input. However, most available data are given in spatially aggregate form, e.g. by zones. Micro data of households and workplaces, residences and businesses are rarely available, and where they are, their use is restricted for privacy reasons.

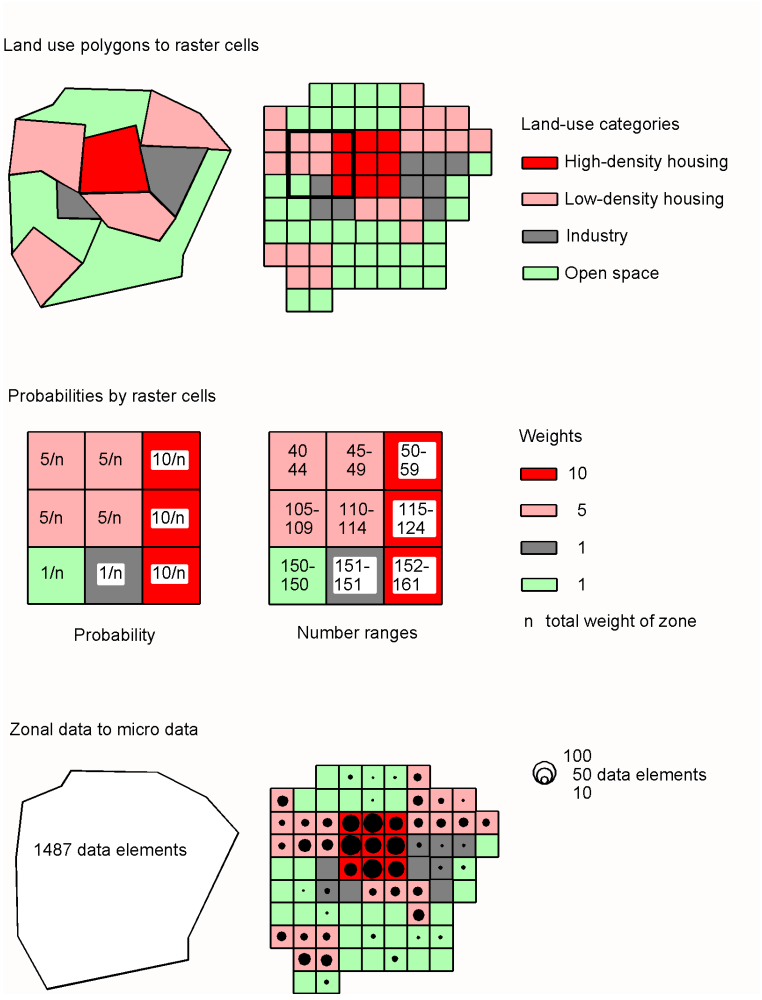

Figure 4 Raster disaggregation of zonal data (Spiekermann and Wegener 2000: 48)

Where no micro data are available, GIS can be used to generate a synthetic disaggregate spatial micro database which corresponds to all known statistical distributions (Bracken and Martin 1995, 1989; Martin and Bracken 1991). To achieve this, raster cells or pixels are used as addresses for the microsimulation (Wegener and Spiekermann 1996). To disaggregate spatially aggregate data within a spatial unit such as an urban district or a census tract, the land use distribution within that zone is taken into consideration, i.e. it is assumed that there are areas of different density within the zone. The spatial disaggregation of zonal data therefore consists of two steps, the generation of a raster representation of land use and the allocation of the data to raster cells. Figure 4 illustrates the procedure for a simple example. The following steps are performed (Spiekermann and Wegener 2000):

- First, the land use coverage and the coverage containing the zone borders are overlaid to get land-use polygons for each zone. Then the polygons are converted to a raster representation by using a point-in-polygon algorithm for the centroids of the raster cells. As a result each cell has two attributes, the land-use category and the zone number of its centroid. These cells represent the addresses for the disaggregation of zonal data and the subsequent microsimulation. The cell size selected depends on the required spatial resolution of the microsimulation.

- The next step merges the land-use data and zonal activity data such as population or employment. First, for each activity to be disaggregated specific weights are assigned to each land-use category. Then all cells are attributed with the weights of their land-use category. Dividing the weight of a cell by the total of the weights of all cells of the zone gives the probability that this cell will be the address of one element of the zonal activity. Cumulating the weights over the cells of a zone one gets a range of numbers associated with each cell. Using a random number generator for each element of the zonal activity, one cell is selected as its address.

The result is a raster representation of the distribution of the activity within the zone.

4. SYNTHETIC POPULATION FOR NETANYA, ISRAEL

In the first application presented, a synthetic population was created for the city of Netanya in Israel. Located in the northern periphery of Tel Aviv at the Mediterranean Sea Netanya has about 159,000 inhabitants.

It was founded in 1928 as an independent resort town. With the growing sprawl of the Tel Aviv region Netanya has become part of its metropolitan area.

Statistically, the city of Netanya is divided into 54 zones or eight regions (Figure 5). Most statistical data were available for the 54 zones in one-dimensional tables. Information had to be disaggregated to micro locations and to multidimensional distributions.

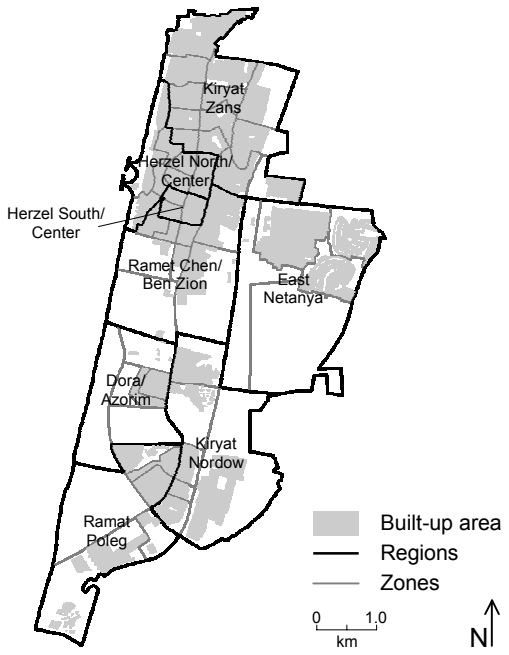

Figure 5 Study area of Netanya

4.1 Procedure

The synthetic population module generated a synthetic population with demographic attributes closely matching the real population of Netanya. Data of administrative registers were transformed by iterative proportional fitting to multidimensional data that served as probabilities to generate a synthetic population by Monte Carlo sampling. The households so generated were then spatially distributed to raster cells.

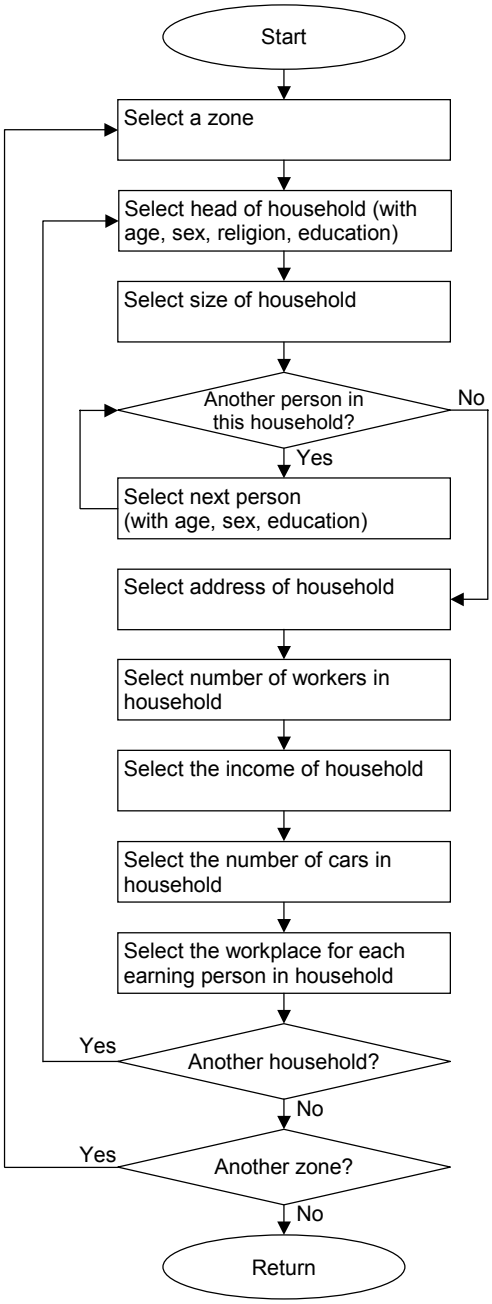

Figure 6 The synthetic population generator

Figure 6 shows the main structure of the synthetic population generator. It was used to generate the population of Netanya with 159,000 persons living in about 50,000 households. To create a household, first the head of household was selected, because many features of a household depend on the characteristics of the head of household. Next, age, gender, religion, and education of the head of household were selected. Depending on this information, the household size was determined. If a one-person-household was selected, no more persons needed to be created for that household. If a multiple-person household was drawn, the other household members were selected in turn until all persons for the household were created. Then the housing location in the zone was chosen. The housing location was a micro location consisting of row and column coordinates of a raster cell of 50 by 50 metres. Aggregate population data were spatially disaggregated to these raster cells using the method described in Section 3.3. Next the number of earning people and then the household income were selected. Depending on all this information, the number of cars was determined. Finally a workplace was assigned to each working person. The programme continued to create households until all households of the zone were created. Then the procedure continued with the next zone until the population for all 54 zones of Netanya was generated.

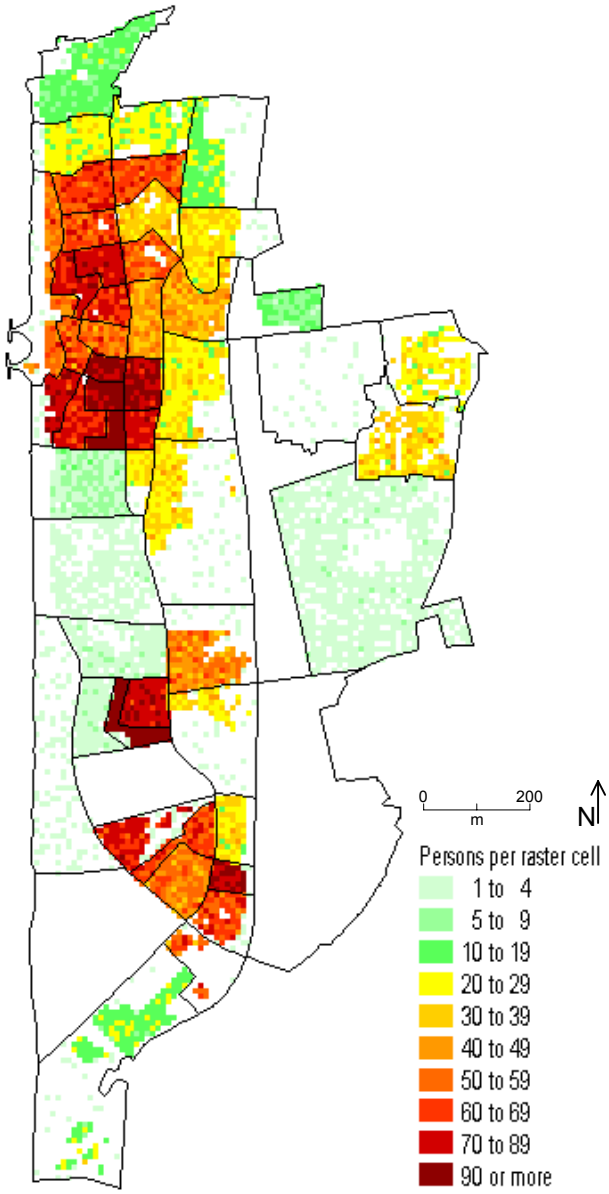

Figure 7 Synthetic population density in Netanya by raster cell

There were some cases where a 'natural order' of influencing the probabilities of following features could not be determined. For instance, on the one hand the number of workers in a household influences how many cars the household can afford. But on the other hand the number of cars might influence the number of workers, since more jobs are accessible by car. In such cases it was assessed which feature is likely to have the stronger impact on the other. In the given example, it was assumed that the influence of the number of workers on the number of cars is stronger than vice versa.

There were some improbable constellations. As the size of a household depends on the age of the head of household, an eight-person household headed by a 20-year-old person is very unlikely. Wherever possible, such cases are excluded. But a small number of exceptions are possible. There are some specific cases where for instance the father left the household and the oldest son takes over as head of household. But generally these constellations were avoided.

4.2 Output

Figure 7 shows the resulting distribution of persons by raster cells. The shading indicates population density, i.e. number of persons per raster cell. The numbers have to be multiplied by four to indicate persons per hectare.

Besides maps like Figure 7, results of the synthetic population generator programme are two lists: a list of households and a list of persons that live in these households. The two files are linked by the corresponding household numbers and person numbers. Figure 8 presents excerpts of the household file <hh.dat>. Each line describes one household. Figure 9 presents excerpts from the person file <pp.dat>. Every line in this file represents one person.

# of household	Zone ID	Column (x-coordinate)	Row (y-coordinate)	Income in NIS	Number of earners	Number of cars	Household size	Person #1	Person #2	Person #3	Person #4	Person #5	Person …
1	111	79	65	6115	2	1	3	1	2	3			
2	111	75	69	4307	0	0	2	4	5				
3	111	75	67	9516	3	1	5	6	7	8	9	10	
4	111	78	65	5215	2	1	4	11	12	13	14		
5	111	77	69	6956	1	0	4	15	16	17	18		
6	111	73	67	10215	3	1	3	19	20	21			
7	111	78	63	6581	1	1	3	22	23	24			
8	111	78	66	8436	0	1	5	25	26	27	28	29	
9	111	81	63	6331	1	1	2	30	31				
10	111	77	65	8564	0	1	5	32	33	34	35	36	
11	111	81	74	8705	2	1	5	37	38	39	40	41	
12	111	85	76	4942	2	1	4	42	43	44	45		
...													

Figure 8 Excerpts from household file <hh.dat>

# of person	# of household	Age	Gender	Religion	Education	Place of work
1	1	44	1	1	1	23
2	1	41	2	1	1	9
3	1	17	2	1	5	
4	2	70	2	1	4	
5	2	73	1	1	3	
6	3	56	2	1	2	8
7	3	58	1	1	1	
8	3	24	2	1	2	23
9	3	23	1	1	3	15
10	3	20	1	1	3	
11	4	66	2	2	7	
12	4	66	1	2	1	
13	4	42	2	2	7	2
14	4	40	1	2	1	
...						

Figure 9 Excerpts from person file <pp.dat>

As the first household in <hh.dat> is a three-person household, it is represented in the person file <pp.dat> by three lines. The first number in each line is the sequential number of the person and the second number is the number of the household. These two numbers correspond to the person number and the household number in <hh.dat>. The next field contains the age of the person, followed by the gender, where 1 stands for male and 2 for female. The next number indicates the religious affiliation of the person: 1 means non ultra-orthodox and 2 ultra-orthodox. The following number indicates the level of education of the person, where seven levels of education are distinguished. The last number indicates the place of work, if any.

5. SYNTHETIC POPULATION FOR DORTMUND, GERMANY

The project ILUMASS (Integrated Land-Use Modelling and Transportation System Simulation) aims at embedding an existing microscopic dynamic simulation model of urban road traffic flows into a comprehensive model system incorporating changes of land use, the resulting changes in transport demand and the environmental impacts. This project requires a synthetic population with households and persons, businesses and public actors, dwellings and non-residential floorspace.

The study region of ILUMASS is the urban region of Dortmund (Figure 10). The area consists of Dortmund and of its 25 surrounding communities. The area is subdivided into 246 statistical zones, and the micro data were generated for every zone. However, even this spatial resolution is not sufficient for analysing environmental impacts such as air quality and traffic noise, which require a much higher resolution. The spatial disaggregation technique presented in Section 3.3 was used to disaggregate the zonal data to raster cells of 100 by 100 m size for the calculation of spatially disaggregate environmental and equity indicators. There are about 200,000 raster cells in the study area. The population is about 2.6 million people living in 1.1 million households. Also about 100,000 businesses had to be generated.

Figure 10 Study region of Dortmund

5.1 Procedure

The basic procedure to generate the micro data is similar to the way the synthetic population was produced for Netanya. There are two main further developments: First, there are by far more features to be generated for every actor. Second, and more importantly, besides households and persons also synthetic populations of businesses, dwellings and non-residential floorspace are generated in order to represent all relevant actors and buildings for land-use transport modelling.

5.2 Output

The synthetic population of ILUMASS represents individual actors in the form of households and household members. Compared to Netanya household are described with more detail. Besides the number of cars, eleven types of car are differentiated, ownership of monthly season ticket and car-sharing membership were generated, and a monthly budget for mobility was determined. In contrast to Netanya, each person is represented by the additional characteristics employment status and ownership of driving license. While the income was represented on a household level in Netanya the income in the Dortmund region is generated for each person separately. Per definition a household shares its income with all household members. However determining the income of single persons gives a better chance to represent different income opportunities depending on age, gender and education.

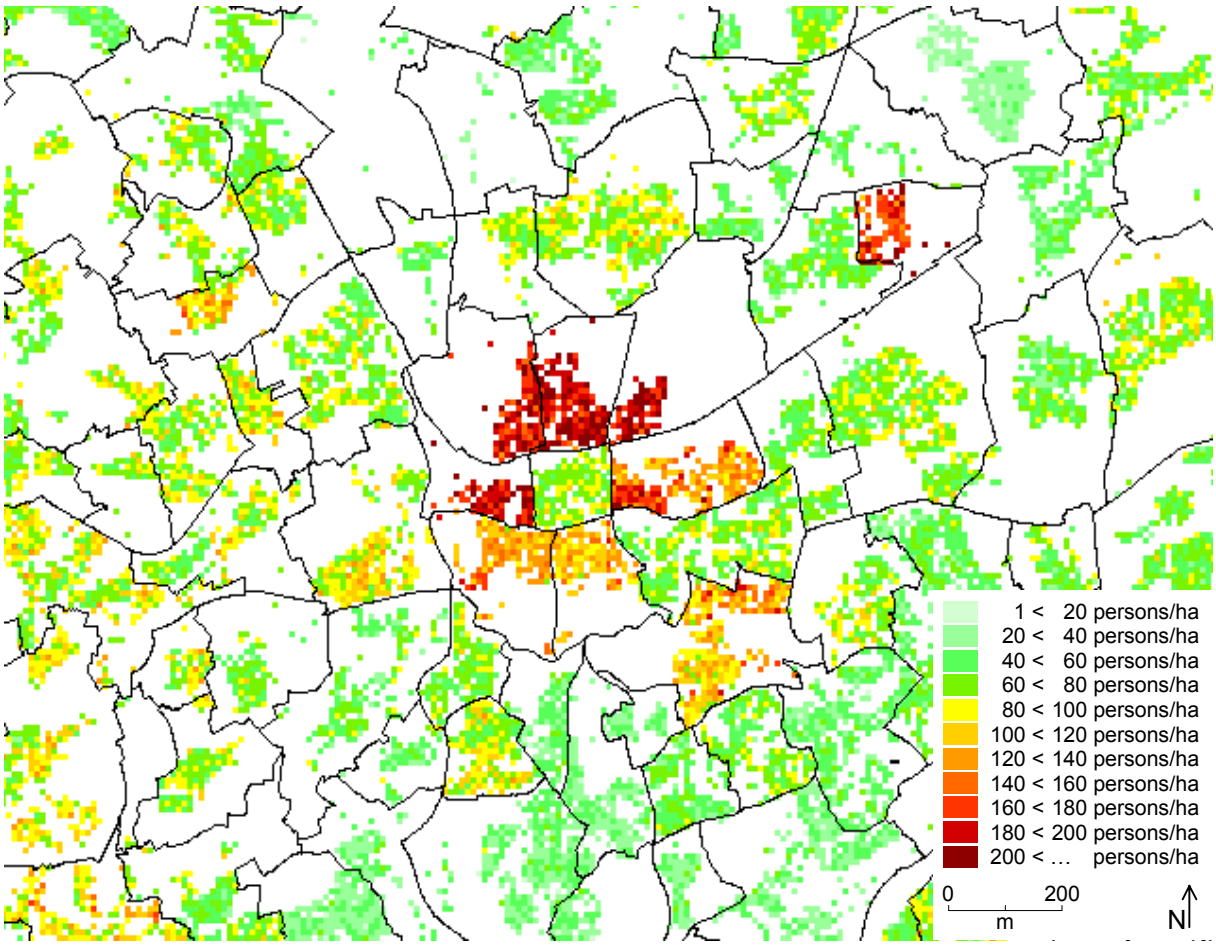

Figure 11 Synthetic population density in the Dortmund study area (excerpt)

Features associated with each dwelling are building type (single family house or apartment house), size, quality, tenure and price. Each dwelling has a raster cell as a micro location given by x and y coordinates in a raster of 100 by 100 metres. Figure 12 shows excerpts from the dwelling file.

# of dwelling	Column (x-coordinate)	Row (y-coordinate)	# of household	Building type	Tenure	Number or rooms	Quality	Monthly costs in Euro
1	216	155	45317	2	2	5	3	547
2	213	156	66132	2	3	2	4	821
3	187	143	11413	2	2	4	4	1421
4	272	173	471647	1	2	5	4	2105
5	364	133	8144	1	1	2	1	840
6	272	173	8973	2	1	2	1	294
7	274	173	1015134	2	3	6	2	927
8	437	145	0	2	3	4	3	486
9	426	132	943141	1	1	6	1	945
10	236	116	249434	2	1	2	1	226

Figure 12 Excerpts from dwelling file

The household number in the dwelling file serves as a link to the corresponding household. Building type distinguishes between single-family house (1), flat (2) and non-residential building (3). The tenure could be owner occupied (1), rented (2) or publicly subsidised (3). The monthly costs are either costs for rent or costs for interests and repayment of loans in Euro.

The synthetic businesses represented the employers in the model. Businesses are described by their industry, number and qualification of employees, number and type of vehicles, capacity to attend customers, and location in the study region. Public facilities are special cases of businesses. They include institutions like kindergartens, schools, universities, hospitals, or museums. Businesses change land use by the establishment, relocation or closure of businesses and affect transport by the person and goods movements they generate.

# of business	industry	Column (x-coorinate)	Row (y-coordinate)	Costs for rent in Euro	Employees				Capacity for customers	Vehicles by type (1)	Vehicles by type (2)	Vehicles by type (3)	Vehicles by type (4)	Vehicles by type (…)
					Qualification 1	Qualification 2	Qualification 3	Qualification 4						
1	58	224	461	8200	12	18	2		12	102	304	805	1106	120
2	2	115	124	16500	68	9	1	1	4426					
3	157	134	345	5900	8	19	43	6	31	103	107			
4	202	48	241	420	2	5	1	1	2	104				
5	55	67	574	600	3		1		152					
6	17	324	37	1860	3	6		1	942	302	405	106		
7	62	412	610	1150	24	2								
8	20	121	8	290			3		18	216				
9	98	34	411	21420	24	238	87	18	10	614	116	5220	4132	3992
10	21	512	42	1820	18	31	2	8	24	101	306	411		

Figure 13 Excerpts from business file

ILUMASS differentiates 65 different kinds of businesses, such as farms, construction companies, shops, banks, hotels, cinemas, youth centres, libraries or parking garages. Figure 13 shows excerpts from the business file. Employees are distinguished by four different levels of qualification. The capacity for customers describes how many people can visit this company at the same time. The number of vehicles are described by four digits. The first and second digit indicate the number of vehicles, the second and fourth digit tell the type of vehicle. This allows to distinguish many different vehicle types.

Businesses are located in non-residential floorspace. For every raster cell of 100 by 100 metres the non-residential floorspace for industry, retail, office, and public use was determined. For each of those four floorspace types the amount of square meters in use, the available area as well as the price are specified. Figure 14 shows excerpts from the non-residential floorspace file.

Column (x-coordinate)	Row (y-coordinate)	Quality	Area in use in m²				Available area in m²				Price per m² in Euro			
			Industry	Retail	Office	public	Industry	Retail	Office	public	Industry	Retail	Office	public
264	154	2	0	600	0	0	500	0	0	0	2	15		
265	154	4	0	0	150	0	0	0	0	0			22	
266	154	1	1300	0	100	0	200	0	0	0	9		24	
267	154	2	0	800	0	0	0	0	0	0		28		
268	154	3	0	0	0	3900	0	0	0	0				8
269	154	3	200	100	700	0	2000	0	0	0	10	21	28	
270	154	3	0	0	0	0	0	400	0	0		18		
271	154	1	0	4600	100	0	0	0	0	0		13	19	
272	154	4	850	0	0	0	0	0	0	0	6			
273	154	2	9100	0	200	0	0	0	600	0	8		24	

Figure 14 Excerpts from the non-residential floorspace file

6. FUTURE WORK

The two applications of generating a synthetic population in Netanya and the Dortmund region show that it is possible to generate synthetic populations as micro data for a microsimulation land-use transport model. Monte Carlo sampling provides a powerful procedure to combine the desired features without restrictions given by input data. The Dortmund synthetic population is used as synthetic data in ILUMASS for the microsimulation of land-use and transport changes and their environmental impact.

Future work will simulate changes of the synthetic populations in each simulation period. So far the events aging, death, birth and divorce/separation have been modelled. Other events include marriage/cohabitation, change of job, retirement, change of income, or obtaining or losing a driving license. Households buy cars and move into, out of or within the study region. Businesses change their location, hire employees or make people redundant, buy cars and commercial vehicles, extend their production or decline. Dwellings and non-residential floorspace are newly constructed, upgraded or demolished. Landlords change the prices of housing in response to demand.

7. REFERENCES

Arentze, T., Timmermans, H. (2000): **ALBATROSS – A Learning Based Transportation Oriented Simulation System.** Eindhoven: European Institute of Retailing and Services Studies.

Barrett, C.L. et al. (1999): **TRansportation ANalysis SIMulation System (TRANSIMS).** Version TRANSIMS-LANL-1.0. Volume 0: Overview. LA-UR 99-1658. Los Alamos National Laboratory, Los Alamos, NM. <http://transims.tsasa.lanl.gov/PDF_Files/Vol0-jmhF_990602_.pdf> [Accessed 18 March 2003]

Batty, M., Conroy, R., Hillier, B., Jiang, B., DeSyllas, J., Mottram, C., Penn, A., Smith, A., Turner, A. (1998): **The virtual Tate.** Working Paper 5. Centre for Advanced Spatial Analysis: London.

Beckman, R.J., Baggerly, K.A., and McKay, M.D. (1996): Creating Synthetic Baseline Populations. **Transportation Research**, Vol. 30, No. 6, 415-429.

Bracken, I., Martin, D. (1995): Linkage of the 1981 and 1991 Censuses using surface modelling concepts. **Environment and Planning A 27**, 379-390.

Bracken, I., Martin, D. (1989): The generation of spatial population distributions from census centroid data. **Environment and Planning A 21**, 537-543.

Chapin, F.S. (1974): **Human Activity Patterns in the City: What People Do in Time and Space.** New York: John Wiley.

Chapin, F.S., Weiss S.F. (1968): A probabilistic model for residential growth. **Transportation Research** 2, 375-390.

Clarke, G.P., ed. (1996): **Microsimulation for Urban and Regional Policy Analysis.** European Research in Regional Science 6. London: Pion.

Clarke, M. (1981): A first-principle approach to modelling socio-economic interdependence using microsimulation. **Computers, Environment and Urban Systems** 6, 211-227.

Clarke, M., Holm, E. (1987): Microsimulation methods in spatial analysis and planning. **Geografiska Annaler** 69B, 145-164.

Clarke, M., Keys, P., Williams, H.C.W.L. (1980): **Micro-Analysis and Simulation of Socio-Economic Systems: Progress and Prospects.** Leeds: School of Geography, University of Leeds.

Deming, W.E., Stephan, F.F. (1940): On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. **The Annals of Mathematical Statistics** 11. 427-444.

Hägerstrand, T. (1970): What about people in regional science? **Papers of the Regional Science Association** 24, 7-21.

Hägerstrand T. (1968): **Innovation Diffusion as Spatial Process.** Chicago: University of Chicago Press.

Hertkorn, G., Wagner, P. (2002): **Generierung einer synthetischen Bevölkerung**. Final Report for SimVV. Ministerium für Schule, Wissenschaft und Forschung des Landes Nordrhein-Westfalen (MSWF).

Holm, E., Lindgren, U., Malmberg, G. (2000): Dynamic microsimulation. In: Fotheringham, A.S., Wegener, M. (Eds.): **Spatial Models and GIS: New Potentials and New Models.** GISDATA 7. London: Taylor & Francis, 143-165.

Hunt, J.D., Donnelly, R., Abraham, J.E., Batten, C., Freedman, J., Hicks, J., Costinett, P.J., Upton, W.J. (2001): Design of a Statewide Land Use Transport Interaction. Model for Oregon. In: WCTR Society (Ed.): **2001 WCTR Proceedings**. Seoul: WCTR Society.

Kain, J.F. and Apgar, W.C. Jr. (1985): **Housing and Neighborhood Dynamics: A Simulation Study.** Cambridge, MA: Harvard University Press.

Kreibich, V. (1979): Modelling car availability, modal split, and trip distribution by Monte Carlo simulation: a short way to integrated models. **Transportation** 8, 153-166.

Landis, J.D., Zhang, M. (1998a): The second generation of the California urban futures model. Part 1: Model logic and theory. **Environment and Planning B: Planning and Design** 25, 657-666.

Landis, J.D., Zhang, M. (1998b): The second generation of the California urban futures model. Part 2: Specification and calibration results of the land-use change submodel. **Environment and Planning B: Planning and Design** 25, 795-824.

Los Alamos National Laboratory (2003): **Population Syntheziser.** LA-UR 00-1725, TRANSIMS 3.0. <http://transims.tsasa.lanl.gov/TRANSIMS_3-0_docs/PDF-V3-0/Ver_3-0_Vol3-Ch2-PopSynth-08Jan03.pdf> [accessed 24 March 2003]

Mackett, R.L. (1990a): **MASTER Model (Micro-Analytical Simulation of Transport, Employment and Residence)**. Report SR 237. Crowthorne, Berkshire: Transport and Road Research Laboratory.

Mackett, R.L. (1990b): Comparative analysis of modelling land-use transport interaction at the micro and macro levels. **Environment and Planning A** 22, 459-75.

Martin, D., Bracken, I. (1991): Techniques for modelling population-related raster databases. **Environment and Planning A 23**, 1069-1075.

Miller, E.J. (2001): **Integrated Land Use, Transportation, Environment (ILUTE) Modelling System.** <http://www.ilute.com/> [accessed 10 March 2003]

Miller, E. (1996): **Microsimulation and Activity-Based Forecasting.** <http://tmip.fhwa.dot.gov/clearinghouse/docs/abtf/miller.stm> [accessed 21 March 2003]

Orcutt, G.H., Greenberger, M., Rivlin, A., Korbel, J. (1961): **Microanalysis of Socioeconomic Systems: A Simulation Study.** New York: Harper and Row.

Salomon, I., Waddell, P., Wegener, M. (2002): Sustainable life styles? Microsimulation of household formation, housing choice and travel behaviour. In: Black, W.R., Nijkamp, P. (Eds.): **Social Change and Sustainable Transport.** Bloomington: Indiana University Press. 125-131.

Schelhorn, T., O'Sullivan, D., Haklay, M., Thurstain-Goodwin, M. (1999): **STREETS: An Agend-Based Pedestrian Model.** Working Paper 9. Centre for Advanced Spatial Analysis: London.

Spiekermann, K., Wegener, M. (2000): Freedom from the tyranny of zones: towards new GIS-based models. In: Fotheringham, A.S., Wegener, M. (Eds.): **Spatial Models and GIS: New Potential and New Models.** GISDATA 7. London: Taylor & Francis. 45-61.

Ton, T., Hensher, D.A. (2001): Synthesising Population Data: The Specification and Generation of Synthetic Households in TRESIS2.0. In: WCTR Society (Ed.): **2001 WCTR Proceedings**. Seoul: WCTR Society.

Torrens, P.M. (2001): **Can Geocomputation save urban simulation? Throw some agents into the mixture, simmer, and wait…** Working Paper 32. Centre for Advanced Spatial Analysis: London.

Veldhuisen, J., Timmermans, H., Kapoen, L. (2000): RAMBLAS: a regional planning model based on the microsimulation of daily activity travel patterns. **Environment and Planning A**, vol. 32, 427-443.

Waddell, P. (2002): UrbanSim: Modeling Urban Development for Land Use, Transportation and Environmental Planning. **Journal of the American Planning Association**, Vol. 68 No. 3, 297-314.

Waddell, P. (2000): A behavioral simulation model for metropolitan policy analysis and planning: residential location and housing market components of UrbanSim. **Environment and Planning B: Planning and Design**, Volume 27. 247-263.

Wegener, M. (1999): **Die Stadt der kurzen Wege: Müssen wir unsere Städte umbauen?** Berichte aus dem Institut für Raumplanung 43. Dortmund: Institut für Raumplanung.

Wegener, M. (1985), 'The Dortmund housing market model: A Monte Carlo simulation of a regional housing market. In: Stahl, K. (Ed.): **Microeconomic Models of Housing Markets.** Lecture Notes in Economic and Mathematical Systems 239. Berlin/Heidelberg/New York: Springer, 144-191.

Wegener, M., Spiekermann, K. (1996): The potential of microsimulation for urban models. In: Clarke, G.P. (Ed.): **Microsimulation for Urban and Regional Policy Analysis.** European Research in Regional Science 6. London: Pion, 147-163.

Wilson, A.G., Pownall, C.E. (1976): A new representation of the urban system for modelling and for the study of micro-level interdependence. **Area**, Vol. 8, 246-254.