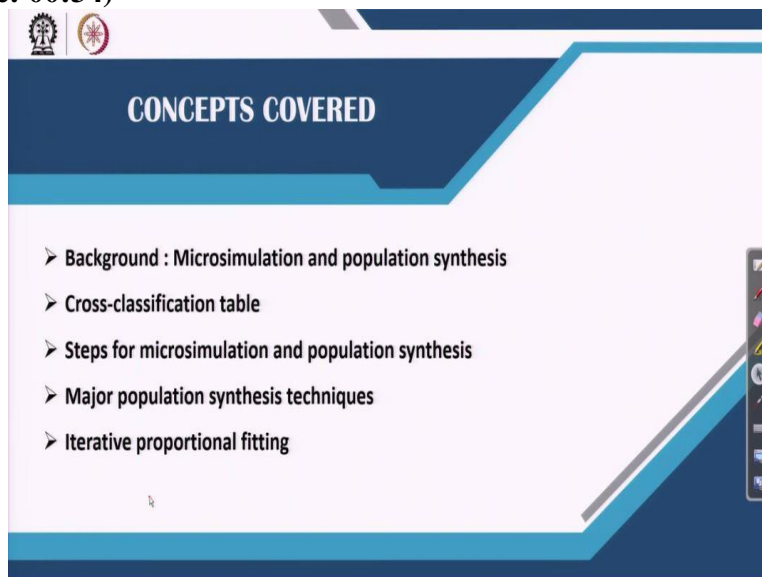


Urban Landuse and Transportation Planning
Prof. Debapratim Pandit
Department of Architecture and Regional Planning
Indian Institute of Technology - Kharagpur

Lecture - 19
Microsimulation and Population Synthesis 1

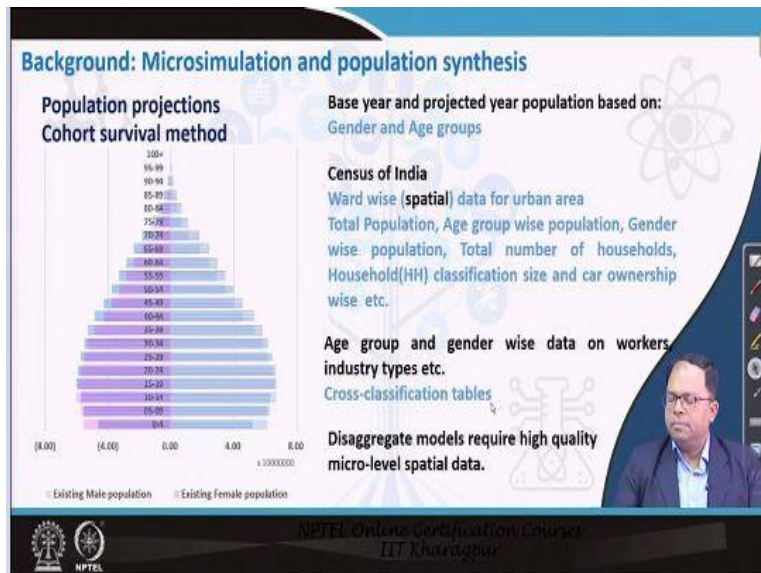
Welcome back to lecture 19. Microsimulation and population synthesis will be covered in two parts - in lecture 19 and lecture 20.

(Refer Slide Time: 00:34)



The different concepts that will be covered in this lecture are on the background of microsimulation and population synthesis, cross-classification tables, different steps for microsimulation and population synthesis, major population synthesis techniques and iterative proportional fitting.

(Refer Slide Time: 00:56)



In the last lecture, different methods of population projections have been covered. During planning of an urban area, there is a need to plan for the current scenario as well as for the future scenario, and for that projected population of the future is required. In the cohort survival method, the population was broken down into different age groups, like in the above figure age group of 0 to 4, 5 to 9 or 10 to 14 and also between male and female population. One side is male and another side is female. Then the population was projected for the next 10 years down the line or maybe 5 years down the line by seeing how each age group transition into the next age group. For example, 0 to 4 people belonging to the 0 to 4 age group, after 5 years will belong to the 5 to 9 age group.

One important thing is that some age groups like around 35 to 39 will also undergo birth and there will be some deaths in the age groups like 85 to 89 and 90 to 94. This will lead to a reduction of the population as well as the addition of the population. Hence, this natural growth is also recorded, and along with that, there is a need to see what are the age groups of the migrant population, how their birth rate or death rate will lead to an increase or decrease of population. These all are measured through the cohort survival method.

In short, in this method, the populations have been broken into different age groups and gender groups and existing residents vs. migrant populations.

The data that is provided in the census of India is based on ward wise. The total population, age group wise population, gender-wise population, total number of households, household classification, the number of people living in that household based on household size and car ownership of those households are given for each ward in an urban area. However, the data on both gender and age group at the ward level may be available or may not be available. But car ownership data for each age group is not available in the census. There are some extra tables in the census e.g., age group and gender-wise data on workers' industry type. Thus, for a particular demographic group, number of workers and in which industry they are working can be figured out. This could be called a cross-classification table where detailed data on each population group or each socio-demographic group in that particular area is available.

In the previous lectures on land use transportation problems, we have seen that the entire land use transportation problem is divided into many smaller components like residential location choice, mobility models, location choice model for businesses and trip generation models, trip distribution models, etc.

Primarily disaggregate models are used to determine or predict the outcome of these different decisions like where will people relocate themselves or which mode a person will undertake. Now, disaggregate models are called disaggregate because they could be used for taking an individual, and then based on his characteristics or his household characteristics or other characteristics, the outcome can be predicted. So, disaggregated models require a high-quality micro level and also spatial data.

This kind of data is not available from the census. But, a lot of micro-level and disaggregated data is required not only for population projection methods but also for other kinds of models.

(Refer Slide Time: 05:32)

Total number of trips made by a family in a particular zone(aggregate estimate):
 HH type(size), Car ownership(0,1,1+)

Cross-classification table for Household in each zone
 Car ownership per HH type: HH(1)Car(0), HH(1)Car(1), HH(1)Car(1+)
 HH(2)Car(0), HH(2)Car(1), HH(2)Car(1+)

Ward 1	HH(1)	HH(2)	HH(3)	HH(4)	HH(4+)	Total(Control totals)
Car(0)	400	600	800	700	500	3000
Car(1)	70	100	120	130	120	540
Car(1+)	0	30	40	35	30	135
Total(Control totals)	470	730	960	865	650	3675

Mode choice for a particular trip in a particular zone(Aggregate/Disaggregate model):
 HH type(size), Car ownership(0,1,1+), Zone characteristics, Modes available, Mode characteristics and Trip type.

Individual: Gender, Age group

- Combine Individual data with household data.
- Predict zone/ward wise modal choice for each individual as per household and trip type.
- Synthetic population based on control totals/marginals.

This kind of cross-classification may not be available in Census data

NPTEL Online Certification Courses
 IIT Kharagpur

in this example (in the above figure), 2 variables (i.e. household type and car ownership) are taken. Household size is used as the categorized result for the household type and for car ownership three categories have been taken like having 0 cars or 1 car or more than 1 car. The total number of trips by a family in a particular zone can be determined based on these 2 particular categories (i.e. household type and car ownership) because the number of trips made by a family depends on the size of that household. Similarly, if the household owns a car, it will be making a certain number of trips, if it does not own a car, it will be making a different number of trips. So, these are influencing factors. Thus, to make an aggregate estimation of the kind of trips based on each of these household types or car ownership, a cross-classification table for households in each zone is required. For each zone, a table will be created where different household sizes and car ownership values will be distributed across rows and columns respectively as shown in the above figure.

In this example, the total number of families having 0 cars is 3000, the total number of families having 1 car is 540 and the total number of families having more than 1 car is 135. Similarly, the total number of household sizes is 470. The total number of household sizes 1 and household size of 2 is 730 and 960 respectively.

Census provides this major data. However, data related to the number of households with size 1 having 0 cars is not available in the census. So, this sort of disaggregated data is required to

determine the mode choice of this particular group of individuals using certain methods. That is why even for an aggregate estimation, this sort of classifications is needed.

The totals of each household type are called control totals because this controls the total number and distribution of households in that entire zone or distribution of houses with different car sizes in the entire zone.

There are several variables which influence mode choice of commuters which can be included in the aggregate or disaggregate mode choice model to predict the mode choice of a particular household or a particular individual in a household such as household type, car ownership, zone characteristics from where the trip is being made, alternative modes available in that particular zone, mode characteristics of those alternative modes, the trip type, individual socio-demographic characteristics (i.e. gender of the individual and the age group of the individual etc.). Hence, for mode choice prediction not only the household data but also individual data needs to be combined with household data, and also there is a need to predict zone or ward-wise modal choice for each individual, as per household and trip type.

As these data are not available, a synthetic population is required to be created based on the control totals or marginals which are available in the census. The synthetic population would have the required detailed data like gender, age group, household type, car ownership etc which could be used for the development of various land-use transportation models like mode choice model, relocation model as well as help us to conduct the entire simulation for this urban area (i.e. microsimulation).

(Refer Slide Time: 10:45)

□ **Microsimulation** is used for predicting the state of an urban area/system by **simulating the behavior of the individual actors**(individuals, households, firms, real estate developer) in the system and in the process access the **impact of current or proposed policies** such as travel demand management(TDM) policies.

□ **Synthetic population** for an urban area is created using **algorithms** which uses **categorical or ordinal socioeconomic data** sets available from secondary sources(Census of India, District statistical handbook) in table format.

One-way, two-way, or multi-way cross tabulations for households, individuals or firms
(Joint aggregate distribution of demographic and/or socio-economic variables)

□ The goal is to create a **list of all households or individuals**(or both in most cases) **or firms** for an urban area and for each zone(ward/grid/TAZ) based on the (Ward/grid/TAZ) **control totals for each attribute category**.

NPTEL Online Certification Courses
IIT Kharagpur

Microsimulation is used for predicting the state of an urban area or system by simulating the behavior of the individual actor, which could be either individuals or households or firms or the real estate developer in the system. This process assesses the impact of current or proposed policies like travel demand management policies.

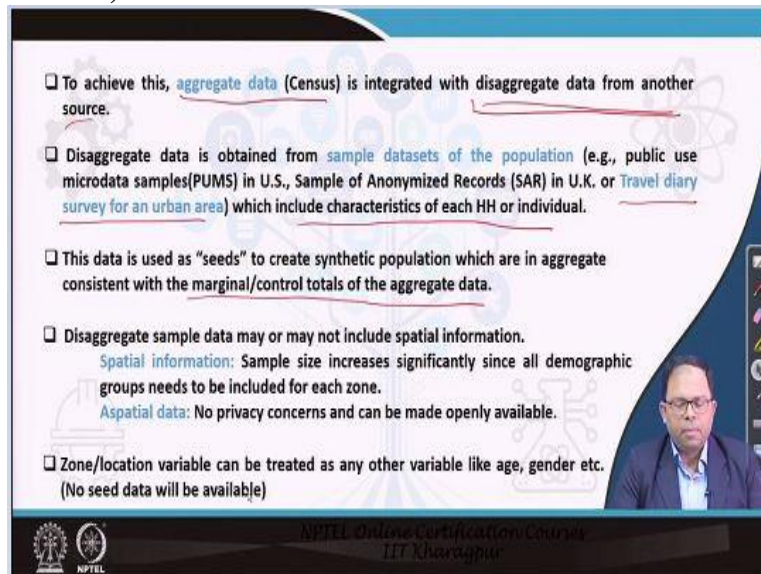
The synthetic population for an urban area is created using algorithms that use categorical, ordinal, socio-economic data sets available from secondary sources like the census of India, district statistical handbook in table format. Based on these available data different cross-tabulation tables can be created. These tables could be a one-way table (i.e. inclusion of only one variable) or a 2-way table (i.e. 2 variables involved) or a multivariate table (i.e. involving many variables and cross dependencies).

For example, if there is gender as well as age group, it could be a 2-way categorization of the entire individuals of a particular area whereas, if gender, age group as well as education are considered then this is 3 dimensions (multi-way cross-tabulations) as a certain number of individuals can be categorized based on gender and education, age group and education, gender and age group.

During the development of multi-way cross-tabulations, the joint distribution can be established. From this tabulation, how many individuals are belonging to each of these groups can be found which can be added further to get the total group and their probabilities can be also developed. It

can be called a joint aggregate distribution of demographic or socio-economic variables. So, this kind of multivariate tables could be used to create a list of all households or individuals or both i.e., if the total number of people of different education groups or age groups or gender groups in each ward or each grid or each TAZ is available, then a list of all households or individuals can be created.

(Refer Slide Time: 14:51)



The slide contains the following text:

- To achieve this, aggregate data (Census) is integrated with disaggregate data from another source.
- Disaggregate data is obtained from sample datasets of the population (e.g., public use microdata samples(PUMS) in U.S., Sample of Anonymized Records (SAR) in U.K. or Travel diary survey for an urban area) which include characteristics of each HH or individual.
- This data is used as "seeds" to create synthetic population which are in aggregate consistent with the marginal/control totals of the aggregate data.
- Disaggregate sample data may or may not include spatial information.
 - Spatial information: Sample size increases significantly since all demographic groups needs to be included for each zone.
 - Aspatial data: No privacy concerns and can be made openly available.
- Zone/location variable can be treated as any other variable like age, gender etc. (No seed data will be available)

The slide also features the NPTEL logo and the text "NPTEL Online Certification Course IIT Kharagpur" at the bottom. A small video inset in the bottom right corner shows a man speaking.

The control totals for smaller zones are available in the census of India which can be integrated with disaggregated data gathered from as a sample from the overall individuals of that particular area. For example, in the US, 'Public Use Micro-data Samples' (PUMS) is used and in the UK 'Sample of Anonymized Records' (SAR) is used. These are surveys conducted by the government, where they collect data from around 1 to 2% of the population i.e., detail data on their age groups, gender, their household properties, or whichever properties the government wants to collect from that particular household or that particular individual. However, in India, these kinds of data sets are not available.

Hence, In India, whenever land-use transportation model is developed for an urban area, data from a large sample of individuals (around 1% or even 2% of the total) have been collected in form of a travel diary survey for an urban area. This could act as the disaggregate sample data set which includes both individuals as well as household characteristics of that particular zone.

This disaggregated data can be used to construct the detailed data sets for each zone for a particular area if the totals of those zones are available from census or similar other data sources. However, this disaggregate data is usually not having any spatial information because of privacy concerns whether it is collected by the government or through a primary survey by various stakeholders (e.g. planner, researchers, policymakers, investors etc.). The census data has got the zoning information but in form of totals. Thus, these two types of data are needed to be combined.

The disaggregated data are used as seeds to create a synthetic population (i.e. the entire list of the population for that particular zone) which are in aggregate consistent with the marginal control totals of the aggregate data (i.e. census data) and this data acts as seeds for each zone for which the synthetic population needs to be generated. Hence, disaggregate sample data may or may not include spatial information to overcome the privacy issue. The only challenge is that large sample size is required so that all demographic groups for each zone are represented.

In this process of synthetic population generation, the zone location variable can be treated like any other variable such as age, gender, etc. which means how many households or individuals are there in each zone is only required to be known.

(Refer Slide Time: 20:04)

Steps for Microsimulation and Population Synthesis

- Step 1: Construct the household-level multi-way table (as per requirement of the microsimulation/individual model)
- Step 2: Construct the individual-level multi-way table
- Step 3: Compute selection probability of an household for a particular zone/grid/TAZ
- Step 4: Select a sample household using Monte Carlo simulation
- Step 5: Appropriateness of the household considering zone control totals for both household and population
- Step 6: Selected household is added to the selected area (Household and population list created)
- Step 7: Household and individual level counts are then updated for each zone
- Step 8: Components of urban land use transportation model (residential mobility, location choice, trip generation, destination choice, mode choice) are simulated using HH and Population list for each zone

Allocation

Fitting

Micro-simulation

NPTEL Online Certification Course
IIT Kharagpur

Steps for Microsimulation and Population Synthesis

There are different methods for population synthesis. The entire process of microsimulation and population synthesis can be divided into 3 parts such as fitting followed by allocation and micro-simulation part.

Fitting

In the fitting part, as per the requirement of the microsimulation or individual models, the household level and individual level multi-way cross-tabulation tables are required to be formulated separately.

Allocation

In the allocation part, how the simulated population or synthetic population for a particular area can be generated, is determined by using the data in the multi-way tables. There are multiple steps to this. In the first step, the selection probability of a household (i.e. sample) for a particular zone or grid or TAZ is computed which could be determined based on the cross-tabulation tables because, in cross-tabulation tables, different demographic groups and exact probability or marginal totals of each of this group is available. Hence, the selection probability of a particular household in a particular zone can be predicted by using conditional probability and after that, a sample can be selected.

A few sample households using Monte Carlo simulation from a list of sample houses can be chosen. For example, data for 10,000 or 5000 people from an urban area is collected, a few of those can be selected, and using those selected samples, the synthetic population of that particular zone can be generated. As the samples cannot be selected randomly it is important to understand their selection probability.

Some samples from the entire data set can be collected by using the Monte Carlo simulation. Then, the appropriateness of that household should be checked by considering zone control totals for both households and population. After that, that selected household can be added to this particular area for which particular synthetic population is being generated and a household and

population list for that area is created and also household and individual level count for that particular zone is updated. The entire process is repeated for the selection of another house.

At a time one household from the sample is selected and put into the target area till the target area is filled while simultaneously checking if the control totals are matching in the target area. Once this control totals are matched then it is recorded i.e. a particular kind of household has been selected from the sample list and thus during the next time the probability of that particular sample being selected can be reduced or kept the same.

Microsimulation stage

After the creation of a list using this allocation strategy of synthetic population, different components of urban land-use transportation models like residential mobility, location choice, trip generation, destination choice, mode choice are simulated using household and population list for each zone which is the part of microsimulation.

In short, microsimulation and population synthesis go hand in hand and the quality of the micro simulation in an urban area depends on the quality of population synthesis.

(Refer Slide Time: 24:50)

Major population synthesis techniques: (Huang and Williamson 2002)

- Iterative proportional fitting and synthetic reconstruction technique
- Combinatorial optimization (CO) technique.

Iterative proportional updating approach
Markov process based approaches
Other emerging approaches

Population synthesizers in use:

PopSynWin: HH selection probabilities are used to control person-level constraints. Chicago, Illinois.

PopGen: Standalone software package for U.S. Fits household and person marginal simultaneously using a novel technique.

ILUTE: Uses a sparse list structure (one entry per unique combination of attributes). Toronto, Canada.

CEMDAP: Combines multiway contingency tables.

ALBATROSS: HH distribution is estimated from the person-level distribution. Europe.

NPTEL Online Certification Courses
IIT Kharagpur

Major population synthesis techniques

Major population synthesis techniques are iterative proportional fitting and synthetic reconstruction techniques (IPFSR). There is also a combinatorial optimization technique (CO). In addition, there are several other techniques like the iterative proportional updating approach, Markov process-based approaches etc. In the rest of the lecture, iterative proportional fitting and combinatorial optimization will be discussed in detail.

Various software are also used like PopSynwin where household selection probabilities are used to control person-level constraints. This is used in Chicago, Illinois. PopGen is a standalone software package for the US which can fit households and personal level marginals simultaneously using a novel technique. There is the ILUTE land-use transportation model where a sparse list structured (one entry per unique combination of attributes) is used and this is used in Toronto, Canada. Then in CEMDAP which is embedded in landuse transportation model and in ALBATROSS model, multi-way contingency tables are combined. This is used in Europe. Here household distribution is estimated from personal level distribution.

(Refer Slide Time: 26:38)

Iterative proportional fitting and synthetic reconstruction (IPFSR)

- The term 'Iterative Proportional fitting' was first introduced in the year 1940 by Deming and Stephan.
- IPFSR is a two stage process: Fitting + Allocation. (Wilson and Pownall, 1976) (Beckman et al., 1996)

Fitting:
Contingency table of attributes of agents sampled from a region is available with marginal totals.
For individual Zones only marginal totals are available.

Using IPF algorithm, a multi-dimensional contingency table or cross-classification table of attributes of households can be obtained for zone as well.

(a) the number of agents in a given category matches the corresponding marginal sum, and
(b) the correlation structure of the seed is retained.

The diagram illustrates the process: A 'Study region' is divided into 'Zone1', 'Zone2', and 'Zone3'. A 'Sampled data contingency table' is derived from the study region, with marginal totals for zones 1, 2, and 3. This table is used as a 'Seed' for the IPF algorithm. The algorithm iteratively adjusts the table until it matches the 'Target' marginal totals. The final output is an 'Estimated contingency table for Zone 1, 2, 3'.

[Source: Choupari et al., 2014]

Iterative proportional fitting

The 'iterative proportional fitting and synthetic reconstruction' (IPFSR) technique has 2 stages which are fitting and allocation. The iterative proportional fitting part was introduced in the year 1940 by Deming, and Stephen. In this process, contingency tables of attributes of agents are sampled from a region with the help of marginal totals.

As for individual zones, only marginal totals are available and using the IPF algorithm a multidimensional contingency table or cross-classification of attributes of households table for every zone has to be created. The number of agents in a given category must match the corresponding marginal sum and the correlation structure of the seed should be retained. For example, the following cross-classification table shows the marginal totals.

Target	Household size				
Car ownership	1 member	2 member	3 member	3 member +	Marginals
0 car					100
1 car					90
1 car +					110
Marginals	90	80	60	70	300

In the next table, the number of people in each of these groups is given which is seed data. Now, this has to be applied for the area where only marginal totals are available as the previous table.

Seed	Household size				
Car ownership	1 member	2 member	3 member	3 member +	
0 car	16	8	12	14	50
1 car	6	12	10	4	32
1 car +	18	22	6	2	48
	40	42	28	20	130

For example, the sample may cover around 10,000 people or 1000 people, whereas this zone may have 1,00,000 people. Hence, there is a need to clone this population for this entire zone using the IPF algorithm. The seed data is collected for the entire study area, but only for 1% of the population or maybe even less. After that, based on this sample seed data category wise population data is generated for each zone. It is assumed that the same correlation structure of the seed will be retained for each of these groups.

(Refer Slide Time: 29:22)

□ For each control variable, the corresponding slice of the contingency table is scaled proportionally so that the total number of agents matches the control total in the target table.

$$P^{k+1}(i,j) = \frac{P^k(i,j)}{\sum_j P^k(i,j)} Q(i)$$

$$P^{k+2}(i,j) = \frac{P^{k+1}(i,j)}{\sum_i P^{k+1}(i,j)} Q(j)$$

Where,
 $P^k(i,j)$ = matrix element in row i, column j, and iteration k.
 $Q(i)$ = Row sum
 $Q(j)$ = Column sum.
 Matrix cell values are estimated iteratively.

(Source: Huang and Williamson, 2001)

□ All control variables are iterated one after another.
 □ This is again repeated and each step is called an iteration.
 □ Iteration is carried on till the relative error of the distribution vs. the marginal sums reaches a user-specified threshold.

NPTEL Online Certification Course
 IIT Kharagpur

In the IPF algorithm for each control variable, the corresponding slice of the contingency table is scaled proportionately. So, the total number of agents matches the control total in the target table. In a contingency table (above table) there are the row values and column values of the control totals and these two also have to be matched.

Hence, for each control variable, the contingency table is modified proportionately so that the total number of agent matches the control total of the target table. If there are several iterations, then for P^{k+1} iteration, where k is the iteration number, ij is the matrix element in row i and column j, then $Q(i)$ is the row sum and $Q(j)$ is the column sum. The matrix and values can be estimated through iterations by using these 2 following formulas

$$P^{k+1}(i,j) = \frac{P^k(i,j)}{\sum_j P^k(i,j)} Q(i)$$

$$P^{k+2}(i,j) = \frac{P^{k+1}(i,j)}{\sum_i P^{k+1}(i,j)} Q(j)$$

Where,

$P^k(i,j)$ = matrix element in row i, column j, and iteration k.

$Q(i)$ = Row sum

$Q(j)$ = Column sum.

Matrix cell values are estimated iteratively.

All control variables are iterated one after another. At first, the row totals are taken care of and then column totals. Then the entire step is again repeated. Each step is called iteration and iteration will be carried on until the relative error of the distribution versus the marginal sums reaches a user-specified threshold. During these iterations, the values of the existing seed table are changed and gradually approach the target values. But, it cannot be achieved in one iteration and it takes a certain number of iterations. The number of iterations are decided based on the amount of the desired accuracy.

(Refer Slide Time: 32:46)

The slide displays the following tables and formulas:

Control total/marginals					
Target	Household size				Marginals
	1 member	2 member	3 member	3 member +	
0 car					100
1 car					50
1 car +					110
Marginals	50	80	60	70	300

Disaggregate data					
Seed	Household size				
	1 member	2 member	3 member	3 member +	
0 car	16	8	12	14	50
1 car	6	12	10	4	32
1 car +	18	22	6	2	48
Marginals	40	42	28	20	130

Row adjustment						
Iteration 1	Household size				Total	Row adjustment factor
	1 member	2 member	3 member	3 member +		
0 car	30	16	24	28	100	1
1 car	16.875	31.75	28.125	11.25	90	2.8225
1 car +	41.25	50.466667	33.75	4.833333	130	2.2566667
Total	90.125	100.186667	65.875	43.833333	300	
Column adjustment factor	0.9883333	0.7966667	0.910159	1.5666667		

$$p^{k+1}(i,j) = \frac{p^k(i,j)}{\sum_j p^k(i,j)} Q(i) \quad \sum_i p^k(i,j) = 50 \quad p^k(i,j) = 16$$

$$p^{k+2}(i,j) = \frac{p^{k+1}(i,j)}{\sum_i p^{k+1}(i,j)} Q(j) \quad \sum_i p^{k+1}(i,j) = 90.125 \quad p^{k+1}(i,j) = 32$$

In the classification table (in the above figure), the seed size data from the sample population of a particular area is presented. In this example, household sizes vary from 1 to more than 3 (i.e. 1 member households, 2 member households, 3 member households, and more than 3 member households) and car ownership values vary from 0 to more than 1 car per household (i.e. 0 car households, 1 car houses, and more than 1 car households).

From the sample survey (top right table in the above figure), it has been found that, 16 number of houses which has got 1 member and 0 cars, 6 number of houses which have got 1 member and 1 car, etc. and the total number is 130 which has to be consistent (table below).

Seed	Household size				
Car ownership	1 member	2 member	3 member	3 member +	
0 car	16	8	12	14	50
1 car	6	12	10	4	32
1 car +	18	22	6	2	48
	40	42	28	20	130

In a particular zone, the marginal is known (top left table in the above figure). There are 100 numbers of 0 car households, 90 single car households, 110 more than 1 car households. Besides, there are 90 single-member families, 80 families with household size 2, 60 families with household size 3, and 70 families with household size 3 which also leads to a total of 300 households (table below).

Target	Household size				
Car ownership	1 member	2 member	3 member	3 member +	Marginals
0 car					100
1 car					90
1 car +					110
Marginals	90	80	60	70	300

In the next step, the previously mentioned IPF formula is used where $Q(i)$ is 100. After the row adjustment (bottom-left table in the above figure), the row totals become the same as the marginal but for the column totals, it is not the same (figure below).

	Row adjustment					
Iteration 1					Total	Row adjustment factor
	32	16	24	28	100	2
	16.875	33.75	28.125	11.25	90	2.8125
	41.25	50.4166667	13.75	4.58333333	110	2.291666667
Total	90.125	100.166667	65.875	43.8333333	300	
Column adjustment factor	0.99861304	0.79866889	0.9108159	1.59695817		

Similarly, in the next step, the totals for the columns are modified with the help of the column adjustment factors (below table). This is end of iteration 1. Now, the column total has matched with marginal totals whereas the row total has changed. Hence, the iteration will be repeated.

Column adjustment					
					Total
	31.9556172	12.7787022	21.8595825	44.7148289	111.3087308
	16.851595	26.9550749	25.6166983	17.96577947	87.38914764
	41.1927878	40.266223	12.5237192	7.319391635	101.3021216
Total	90	80	60	70	300

(Refer Slide Time: 36:24)

The screenshot displays a series of iterative adjustment tables for a cross-classification table. The tables are organized into four iterations, each showing row and column adjustment factors and their corresponding values. The 'Cross classification table' and 'Household size' table are also visible.

Cross classification table		Household size				
Car ownership		1 member	2 member	3 member	3 member +	Total
0 car	28	11	19	42	100	
1 car	17	27	27	20	91	
1 car +	45	43	14	9	111	
Total	90	81	60	71	302	

In step 2, step 3 step 4 (above figure) the iterations are repeated similarly until the difference between the estimated marginal totals and the actual marginal total reaches an acceptable limit. In this particular example, the iteration has been stopped after 4 iterations, though it can be further continued even till 7 or 8 or 9 iterations to get more accurate results.

Instead of multiplying the proportionate value with the actual value from the seed, it can be kept in proportional form as well which represents the proportion of that particular group within that total zone. Finally, these values are converted to an integer which may create a discrepancy like 27.9 becomes 28, 17.1 becomes 17. In this particular example, this may not look too much, but when lots of categorizations and dealing with small groups (e.g. having only 3 or 4 households), it may become an important issue. So, integration has to be done very carefully so that overestimation of certain groups in a particular population can be avoided.

(Refer Slide Time: 38:36)

Marginals can be single or multi-dimensional.
 2 dimensional marginal for attribute M and N can also be treated as a single dimensional marginal $(M \times N)$ of its Cartesian product. (Similarly higher dimensions can be converted)
 Similarly, multi-dimensional control totals can be easily converted into single dimensional control totals with more categories.

IPF can estimate only one level of aggregation.
 Agent-level(individual) or group-level(household) attributes but not for both simultaneously.
 One strategy could be to convert all agent-level attributes into group-level attributes and then apply IPF at the group level. (Arentze et al., 2007).

Zone-by Zone approach
 IPF could be run zone by zone(disaggregate sample data for all zones).

Multi-zone approach
 Here agents/groups can be estimated for all zones simultaneously by adding a control total for each zone.
 Since zone wise seed data is not available we fill the seed data with 1.

NPTEL Online Certification Courses
 IIT Khariipur

Marginals could be a single dimension or multi-dimensional. Marginals can be single that means there are total values for a single factor. In the case of 2 way marginal, 2 factors have to be matched. 2-dimensional marginals for attribute M and N can also be treated as a single-dimensional marginal.

Similarly, higher dimension marginals can also be converted. For example, if there are age groups and car ownership, a new group (e.g. an individual with age group and car ownership) can be formed where M into N together becomes one group and if the seed data for that is available, then that seed data can be used for conversion of higher dimensional marginals into single-dimensional marginals. However, for conversion, those kinds of seeds in the data should be available for sample data as well. Similarly, multi-dimensional control totals can be easily converted into single-dimensional control total with more categories.

IPF can estimate only one level of aggregation i.e., it can estimate at individual or group level(household) but both cannot be done simultaneously. Thus, the agents (i.e. individuals) of the zone can be matched with the control totals or the households in a zone can be matched with the control totals, both cannot be taken together. One strategy could be to convert all agent level attributes into group-level attributes and then apply IPF at the group level.

Another issue is that, since IPF can be done zone by zone, there is a requirement for the disaggregate sample data to have seed data for each zone. It can also be done using a multi-zone

approach, which means that, first agent groups can be estimated for all zones simultaneously by adding a control total for each zone and then it can be determined for each zone. That means, first the IPF will be used to determine demographic groups for the entire urban area and then taking zone as 1 level, this population can be divided into different zones. As in most cases, zone-wise seed data is not available, a matrix is used with each cell having 1 value as the seed data, and using the iterative proportional fitting, the number of population in each zone can be found.

(Refer Slide Time: 42:51)

Target	M X N									Seed	M X N							
Zone	11	12	21	22	31	32	Total	Row Adj.	Zone	11	12	21	22	31	32	Total		
1	5	3	9	2	8	13	51	1.275	1	5	3	9	2	8	13	40		
2	4	6	11	5	4	8	44	1.157895	2	4	6	11	5	4	8	38		
3	6	5	3	7	12	2	41	1.171429	3	6	5	3	7	12	2	35		
4	7	2	1	3	6	11	39	1.3	4	7	2	1	3	6	11	30		
Total	30	21	28	21	32	43	175		Total	22	16	24	17	30	34	143		
Iteration 1																		
Row adjustment								Row Adj.	Column adjustment									
6.4	3.8	11.5	2.6	10.2	16.6	51	1.008135	7.0	4.2	11.1	2.6	8.9	16.8	50.58846				
4.6	6.9	12.7	5.8	4.6	9.3	44	0.99192	5.1	7.6	12.3	5.9	4.0	9.4	44.3584				
7.0	5.9	3.5	8.2	14.1	2.3	41	1.009501	7.8	6.4	3.4	8.4	12.3	2.4	40.61413				
9.1	2.6	1.3	3.9	7.8	14.3	39	0.988869	10.1	2.8	1.3	4.0	6.8	14.5	39.43901				
27.13515	19.22951	29.02613	20.43947	36.68872	42.48102	175		30	21	28	21	32	43	135.561				
Col. Adj.	1.105577	1.092071	0.964648	1.027424	0.872203	1.012217												
Iteration 2																		
Row adjustment								Row Adj.	Column adjustment									
7.1	4.2	11.2	2.6	9.0	16.9	51	0.999836	7.1	4.2	11.2	2.6	8.9	16.9	51.00838				
5.1	7.5	12.2	5.9	4.0	9.3	44	0.999919	5.1	7.5	12.2	5.9	4.0	9.3	44.00355				
7.8	6.5	3.4	8.5	12.4	2.4	41	1.000637	7.9	6.5	3.4	8.5	12.3	2.4	40.97391				
9.9	2.8	1.2	4.0	6.7	14.3	39	0.999637	10.0	2.8	1.2	4.0	6.7	14.3	39.01417				
29.9778	21.00185	28.00903	21.00869	32.08049	42.92214	175		30	21	28	21	32	43	135.9858				
Col. Adj.	1.00074	0.999912	0.999678	0.999586	0.997491	1.001814												

Zone-by Zone approach

Zone by Zone approach

In zone by zone approach, if the seed data for each zone is available (like in the above figure), using marginal data, control total, seed data, data of the target zone can be estimated by applying the IPF procedure. For example (in the above figure), there are 3 categories of M and 2 categories of N. Hence, groups are formed like 11 12, then 21 22, and 31 32. Suppose, there are 4 zones. The zone can be also taken as a particular category like the number of cars, or age groups.

(Refer Slide Time: 44:14)

Target Zone	M X N						Total	Seed	Zone	M X N						Total
	11	12	21	22	31	32			11	12	21	22	31	32		
1	1	1	1	1	1	1	51	1.275	1	5	3	9	2	8	13	40
2	1	1	1	1	1	1	44	1.157895	2	4	6	11	5	4	8	38
3	1	1	1	1	1	1	41	1.171429	3	6	5	3	7	12	2	35
4	1	1	1	1	1	1	39	1.3	4	7	2	1	3	6	11	30
Total	30	21	28	21	32	43	175		Total	22	16	24	17	30	34	143

Iteration 1							Row Adj.	Column adjustment
Row adjustment								
1.3	1.3	1.3	1.3	1.3	1.3	1.3	7.65	0.168148
1.2	1.2	1.2	1.2	1.2	1.2	1.2	6.947368	0.168148
1.2	1.2	1.2	1.2	1.2	1.2	1.2	7.028571	0.168148
1.3	1.3	1.3	1.3	1.3	1.3	1.3	7.8	0.168148
4.904323	4.904323	4.904323	4.904323	4.904323	4.904323	4.904323	29.42594	
Col. Adj.	6.117052	4.281936	5.709248	4.281936	6.524855	8.767774		

Iteration 2							Row Adj.	Column adjustment
Row adjustment								
1.3	0.9	1.2	0.9	1.4	1.9	1.9	7.65	0.168148
1.2	0.8	1.1	0.8	1.3	1.7	1.7	6.947368	0.168148
1.2	0.8	1.1	0.8	1.3	1.7	1.7	7.028571	0.168148
1.3	0.9	1.2	0.9	1.4	1.9	1.9	7.8	0.168148
5.044447	3.531113	4.70815	3.531113	5.380743	7.230374	7.230374	29.42594	
Col. Adj.	5.947134	5.947134	5.947134	5.947134	5.947134	5.947134		

Multi-zone approach



Multi-zone approach

Multi-zone approach is applied where zone wise seed data is not available. The total seed data for this sample population is available, though. After making the initial data value as one, the same IPF procedure (like zone by zone approach) is followed. The proportion of people or households in each demographic group can be determined.

(Refer Slide Time: 45:01)

REFERENCES

- Abdoul-Ahad Choupani and Amir Reza Mamdoohi. Population synthesis using iterative proportional fitting (IPF): A review and future research, *Transportation Research Procedia* 17 (2016), 223 – 23.
- Beckman, R.J., Baggerly, K.A., McKay, M.D. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice* 415-429.
- Guo, J. Y., & Bhat, C. R. (2007). Population Synthesis for Microsimulating Travel Behavior. *Transportation Research Record*, 2014(1), 92-100.
- Z. Huang, P. Williamson, A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata, Department of Geography, University of Liverpool (2001)
- Ryan, J., Maoh, H., Kanaroglou, Population synthesis: Comparing the major techniques using a small, complete population of firms (2009). *Geographical Analysis*, 41 (2), pp. 181-203.
- Nik Lomax & Paul Norman (2016) Estimating Population Attribute Values in a Table: "Get Me Started in" Iterative Proportional Fitting, *The Professional Geographer*, 68:3, 451-461.
- Kirill, M., & Kay, W. A. (2010). Population synthesis for microsimulation: State of the art. *SRTC* 2010.

(Refer Slide Time: 45:10)

CONCLUSION

- Disaggregate models require high quality and highly categorized spatial data which are not available directly from standard data sources like Census of India.
- Sample surveys are conducted to collect data at a desired level of categorization which can be used as a seed to generate synthetic data for other zones or regions.
- Population synthesis is conducted in two steps, namely, fitting and allocation.
- Iterative proportional fitting is the most popular fitting technique adopted in most population synthesis software.

Conclusion

Disaggregate models require high quality and highly categorized spatial data, which are not available directly from standard data sources like the census of India. Sample surveys are conducted to collect data at the desired level of categorization which can be used as a seed to generate synthetic data for other zones or regions.

Population synthesis is conducted in 2 steps such as fitting and allocation. Fitting is discussed in this particular lecture and allocation will be taken up in the next lecture. The iterative proportional fitting is the most popular fitting technique which has been adopted in most population synthesis software.

Thank you.