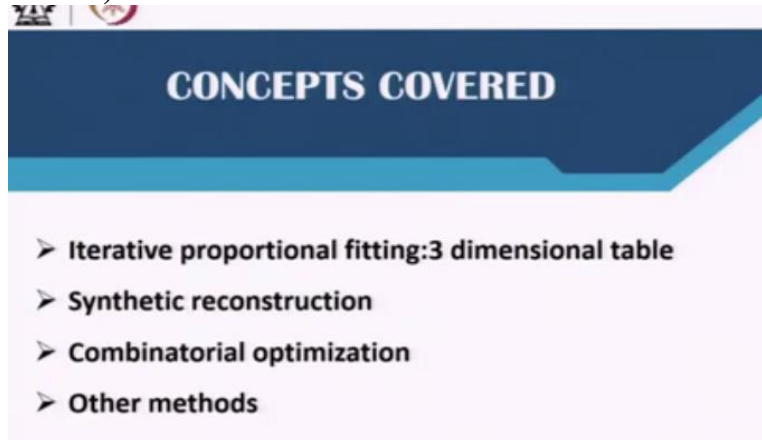


Urban Landuse and Transportation Planning
Prof. Debapratim Pandit
Department of Architecture and Regional Planning
Indian Institute of Science - Kharagpur

Lecture - 20
Microsimulation and Population Synthesis 2

(Refer Slide Time: 00:28)

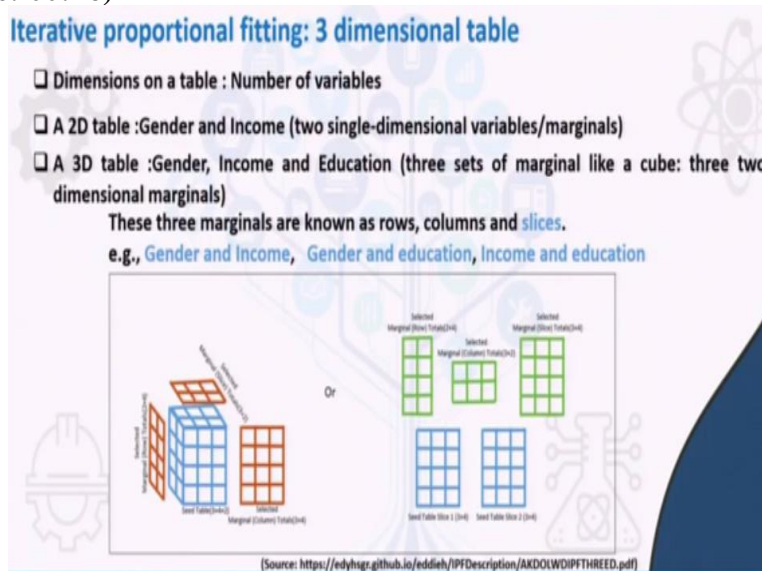


CONCEPTS COVERED

- Iterative proportional fitting: 3 dimensional table
- Synthetic reconstruction
- Combinatorial optimization
- Other methods

Welcome back to lecture 20. In this lecture, the second part of the microsimulation and population synthesis will be discussed which include the proportional fitting for a 3-dimensional table, synthetic reconstruction, combinatorial optimization technique and other methods for generating a synthetic population.

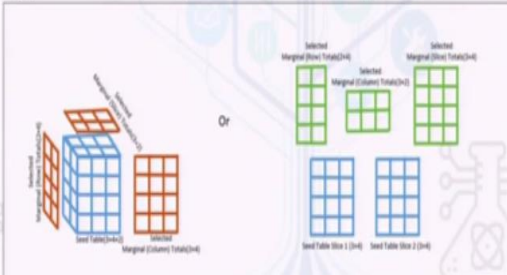
(Refer Slide Time: 00:46)



Iterative proportional fitting: 3 dimensional table

- ❑ Dimensions on a table : Number of variables
- ❑ A 2D table :Gender and Income (two single-dimensional variables/marginals)
- ❑ A 3D table :Gender, Income and Education (three sets of marginal like a cube: three two-dimensional marginals)

These three marginals are known as rows, columns and slices.
e.g., Gender and Income, Gender and education, Income and education



(Source: <https://edyhgr.github.io/eddieh/IPFDescription/AKDOLWDPFTHREED.pdf>)

Iterative proportional fitting: 3-dimensional table

When one additional dimension (i.e. variable) is added with a 2-dimensional table (discussed in lecture 19), it is called a 3-dimensional table. In a 2D table, there are 2 single-dimensional variables (e.g. gender and income) or marginal i.e. the marginals are of a single dimension, but in a 3-dimensional table, there are more marginals or variables (e.g. gender, income, and education) and 2 dimensional marginals (e.g. gender and income, gender and education, and income and education) like a cube.

The seed needs to be 3 dimensional which could be referred to as rows, columns, and also slices. In the above figure, it has been shown how the seed data can be represented as a cube. After taking the cube apart, marginals for rows, columns, and slices can be found along with 2 seed tables for 2 slices.

(Refer Slide Time: 02:54)

- All marginals should add up to the same value.
- The seed should be three-dimensional (e.g. Gender, Income and Education).

Iteration 1

Step 1: Adjust each row of cells (Income-Gender adjustment) proportionally to match the target totals of Marginal 1.

Step 2: Adjust each column of cells (Income-Education adjustment) proportionally to match the target totals of Marginal 2.

Steps 3: Adjust each slice of cells (Gender-Education adjustment) to match the target totals of Marginal 3.

Steps 4 : Repeat Steps 1-3 till convergence is reached.

SEED						
Income	Gender	Slice		Education		
Row	Column	1	2	3	4	Total
1	1	5	4	6	7	22
1	2	3	6	5	2	16
2	1	9	11	3	1	24
2	2	2	5	7	3	17
3	1	8	4	12	6	30
3	2	13	8	2	11	34
Total		40	38	35	30	143
Income						
1		8	10	11	9	38
2		11	16	10	4	41
3		21	12	14	17	64
Gender						
	1	22	19	21	14	76
	2	18	19	14	16	67

Blue: 3 dimensional seed.

Grey: Marginal totals for 2 dimensions

Yellow: Marginal totals for single dimensions

(Source: <http://www.real-statistics.com/matrices-and-iterative-procedures/iterative-proportional-fitting-procedure-ipfp/>)

The same data have been represented in a simpler way (in the figure) instead of a cube form where the blue part is the 3-dimensional seed, the grey part shows the marginal totals for 2 dimensions and the yellow part shows the marginal totals for single dimensions (e.g. 40 is the total number of people of education group 1, 38 is the total number of people have education group 2, similarly, 38 is the total number of people income group 1 and 76 is the total number of people in gender group 1). All marginals should be consistent because they should add up to the total value (e.g. 143).

SEED						
Income	Gender		Slice	Education		
Row	Column	1	2	3	4	Total
1	1	5	4	6	7	22
1	2	3	6	5	2	16
2	1	9	11	3	1	24
2	2	2	5	7	3	17
3	1	8	4	12	6	30
3	2	13	8	2	11	34
	Total	40	38	35	30	143
Income						
1		8	10	11	9	38
2		11	16	10	4	41
3		21	12	14	17	64
	Gender					
	1	22	19	21	14	76
	2	18	19	14	16	67

Now, these marginals (e.g. marginal for income, gender, and education) for a target area must be known for the prediction of the number of households in each of these groups in that particular zone. Each marginal is adjusted proportionately one at a time in the IPF method i.e. this seed data has to be proportionately adjusted as per each marginal. In a first step, each row of cells should be adjusted proportionally to match the target totals of marginal 1 (e.g. income gender adjustment). In the next step, each column of cells should be adjusted proportionately to match the target totals of marginal 2 (e.g. income education adjustment). Similarly, each slice of a cell should be adjusted (e.g. gender education adjustment) to match the target total of marginal 3. Then, steps 1 to 3 will be repeated till convergence is reached like IPF of 2 dimensions (discussed in lecture 19).

(Refer Slide Time: 08:01)

TARGET						SEED											
Income	Gender	Slice		Education		Income	Gender	Slice		Education							
Row	Column	1	2	3	4	Row	Column	1	2	3	4						
1	1					1	1	5	4	6	7						
1	2					1	2	3	6	5	2						
2	1					2	1	9	11	3	1						
2	2					2	2	2	5	7	3						
3	1					3	1	8	4	12	6						
3	2					3	2	13	8	2	11						
Total		51	44	41	39	Total		40	38	35	30						
Iteration 1						Iteration 1											
ROW COLUMN ADJUSTMENT						ROW SLICE ADJUSTMENT						COLUMN SLICE ADJUSTMENT					
Income	Gender	Slice		Education		Income	Gender	Slice		Education		Income	Gender	Slice		Education	
Row	Column	1	2	3	4	Row	Column	1	2	3	4	Row	Column	1	2	3	4
1	1	6.81818	5.45455	8.18182	9.54545	1	1	7.60697	5.7289	6.65896	9.41176	1	1	7.62053	6.90019	5.8842898	9.20091
1	2	3.9375	7.875	6.5625	2.625	1	2	4.39303	8.2711	5.34104	2.58824	1	2	4.38353	6.87142	6.1977782	2.6349
2	1	10.5	12.8333	3.5	1.16667	2	1	11.1429	9.45126	3.745763	1.67606	2	1	12.1645	11.3836	3.3099996	1.63851
2	2	2.47059	6.17647	8.64706	3.70588	2	2	2.85714	4.54874	9.254237	5.32394	2	2	2.85097	3.77897	10.738677	5.41992
3	1	8.53333	4.26667	12.8	6.4	3	1	8.20036	4.74591	13.35994	6.30177	3	1	8.21497	5.71622	11.805711	6.16058
3	2	16.4412	10.1176	2.52941	13.9118	3	2	15.7996	11.2541	2.640061	13.6982	3	2	15.7655	9.34961	3.0635445	13.9452
Total		48.7008	46.7237	42.2208	37.3548	Total		51	44	41	39	Total		51	44	41	39
Iteration 1						Iteration 1						Iteration 1					
Income	Gender	Slice		Education		Income	Gender	Slice		Education		Income	Gender	Slice		Education	
Row	Column	1	2	3	4	Row	Column	1	2	3	4	Row	Column	1	2	3	4
1	1	10.7557	13.3295	14.7443	12.1705	1	1	12.0041	13.7716	12.082068	11.8358	1	1	12.0041	13.7716	12.082068	11.8358
1	2	12.9706	19.0098	12.1471	4.87255	1	2	15.0155	15.1626	14.048677	7.05843	1	2	15.0155	15.1626	14.048677	7.05843
2	1	24.9745	14.3843	15.3294	20.3118	2	1	23.9805	15.0658	14.869255	20.1058	2	1	23.9805	15.0658	14.869255	20.1058
Total		51	44	41	39	Total		51	44	41	39	Total		51	44	41	39
Iteration 1						Iteration 1						Iteration 1					
Income	Gender	Slice		Education		Income	Gender	Slice		Education		Income	Gender	Slice		Education	
Row	Column	1	2	3	4	Row	Column	1	2	3	4	Row	Column	1	2	3	4
1	1	25.8515	22.5545	24.4818	17.1121	1	1	27.9502	19.9261	23.76466	17.3896	1	1	28	24	21	17
1	2	22.8493	24.1691	17.739	20.2426	1	2	23.0498	24.0739	17.23534	21.6104	1	2	23	20	20	22
Total		51	44	41	39	Total		51	44	41	39	Total		51	44	41	39

In the table below, (continuation of the earlier example) marginals (i.e. marginals for income and gender, marginals for gender and education, marginals for income and education) for the target area are available.

TARGET						SEED					
Income	Gender	Slice		Education		Income	Gender	Slice		Education	
Row	Column	1	2	3	4	Row	Column	1	2	3	4
1	1					1	1	5	4	6	7
1	2					1	2	3	6	5	2
2	1					2	1	9	11	3	1
2	2					2	2	2	5	7	3
3	1					3	1	8	4	12	6
3	2					3	2	13	8	2	11
Total		51	44	41	39	Total		40	38	35	30
Iteration 1						Iteration 1					
Income	Gender	Slice		Education		Income	Gender	Slice		Education	
Row	Column	1	2	3	4	Row	Column	1	2	3	4
1	1	12	14	12	12	1	1	8	10	11	9
1	2	15	14	13	7	1	2	11	16	10	4
2	1	24	16	16	20	2	1	21	12	14	17
Total		51	44	41	39	Total		40	38	35	30
Iteration 1						Iteration 1					
Income	Gender	Slice		Education		Income	Gender	Slice		Education	
Row	Column	1	2	3	4	Row	Column	1	2	3	4
1	1	28	24	21	17	1	1	22	19	21	14
1	2	23	20	20	22	1	2	18	19	14	16
Total		51	44	41	39	Total		40	38	35	30

in the first step, for example, the target value 30 which is the summation of the first row, is divided with seed table marginal total to get the proportion value i.e. 30 divided by 22 which is multiplied with 5 resulted in 6.818. The process is repeated for all the rows to get an updated table (shown below), where all the first marginal matches with the target, but none of the other matches.

Iteration 1						
ROW COLUMN ADJUSTMENT						
Income	Gender		Slice	Education		
Row	Column	1	2	3	4	Total
1	1	6.81818	5.45455	8.18182	9.54545	30
1	2	3.9375	7.875	6.5625	2.625	21
2	1	10.5	12.8333	3.5	1.16667	28
2	2	2.47059	6.17647	8.64706	3.70588	21
3	1	8.53333	4.26667	12.8	6.4	32
3	2	16.4412	10.1176	2.52941	13.9118	43
	Total	48.7008	46.7237	42.2208	37.3548	175
Income						
1		10.7557	13.3295	14.7443	12.1705	51
2		12.9706	19.0098	12.1471	4.87255	49
3		24.9745	14.3843	15.3294	20.3118	75
Gender						
1		25.8515	22.5545	24.4818	17.1121	90
2		22.8493	24.1691	17.739	20.2426	85

Hence, similarly, in the next step, the next marginal have been adjusted. For example, for income level 1, the proportion factor would be 12 divided by 10, where 12 is target marginal. Then, the proportion value 1.2 would be multiplied with 6.81818 and 3.9375 resulted in 7.60697 and 4.3903 respectively. A similar process will be repeated for other marginal to get the following updated table (Row slice adjustment table). Now, a similar process is repeated for the column adjustment to complete the first iteration.

ROW SLICE ADJUSTMENT						
Income	Gender		Slice	Education		
Row	Column	1	2	3	4	Total
1	1	7.60697	5.7289	6.65896	9.41176	29.4066
1	2	4.39303	8.2711	5.34104	2.58824	20.5934
2	1	12.1429	9.45126	3.745763	1.67606	27.0159
2	2	2.85714	4.54874	9.254237	5.32394	21.9841
3	1	8.20036	4.74591	13.35994	6.30177	32.608
3	2	15.7996	11.2541	2.640061	13.6982	43.392
	Total	51	44	41	39	175
Income						
1		12	14	12	12	50
2		15	14	13	7	49
3		24	16	16	20	76
Gender						
1		27.9502	19.9261	23.76466	17.3896	89.0305
2		23.0498	24.0739	17.23534	21.6104	85.9695

COLUMN SLICE ADJUSTMENT						
Income	Gender		Slice	Education		
Row	Column	1	2	3	4	Total
1	1	7.62053	6.90019	5.8842898	9.20091	29.6059
1	2	4.38353	6.87142	6.1977782	2.6349	20.0876
2	1	12.1645	11.3836	3.3099996	1.63851	28.4966
2	2	2.85097	3.77897	10.738677	5.41992	22.7885
3	1	8.21497	5.71622	11.805711	6.16058	31.8975
3	2	15.7655	9.34961	3.0635445	13.9452	42.1238
	Total	51	44	41	39	175
Income						
1		12.0041	13.7716	12.082068	11.8358	49.6935
2		15.0155	15.1626	14.048677	7.05843	51.2851
3		23.9805	15.0658	14.869255	20.1058	74.0213
Gender						
1		28	24	21	17	90
2		23	20	20	22	85

(Refer Slide Time: 11:34)

Zero cell problem

- ❑ Not all categories are represented in the seed data. The zero value in the seed of this category leads to problems due to division by zero.

Solution 1: Replace zero cells with a small value (.001)

Solution 2: Category reduction (PopSyn, CEMDAP (pre-processing step))

Categories can be automatically aggregated considering a user specified threshold for marginals.

Category reduction can be an independent step to improve results.

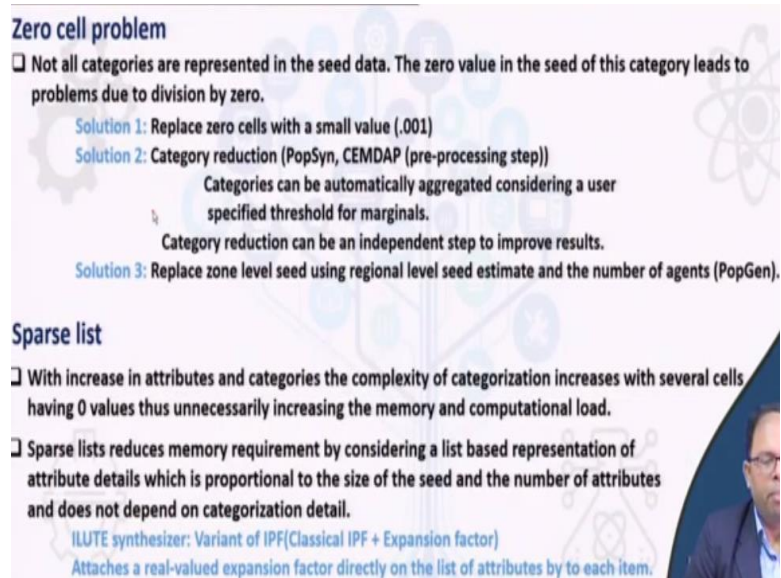
Solution 3: Replace zone level seed using regional level seed estimate and the number of agents (PopGen).

Sparse list

- ❑ With increase in attributes and categories the complexity of categorization increases with several cells having 0 values thus unnecessarily increasing the memory and computational load.
- ❑ Sparse lists reduces memory requirement by considering a list based representation of attribute details which is proportional to the size of the seed and the number of attributes and does not depend on categorization detail.

ILUTE synthesizer: Variant of IPF(Classical IPF + Expansion factor)

Attaches a real-valued expansion factor directly on the list of attributes by to each item.



Zero Cell Problem

One of the major problems with IPF is called the 0 cell problem. In the previous example, all the cells have got some values, but some of these cells could have 0 value, because in the actual case when different demographic groups (not only income gender education) are considered, the total number of people in each group may be very less which results in cells with no value. The '0 value' in the seed leads to problems due to division by 0. There are 3 ways to address this problem.

- the 0 cells can be replaced with a small value which can be like 0.001
- category reduction can be done as in PopSyn and CEMDAP. In this process, before starting the IPF, categories can be automatically aggregated considering a user-specified threshold for marginals. Category reduction could be also done as an independent step to improve results as sometimes too much disaggregation does not provide any additional advantages.
- Zone level seed can be replaced with regional level seed estimate and for that average values from the regional area can be taken (e.g. PopSyn)

Sparse List

With increasing attributes and categories, the complexity of categorization increases with several cells having 0 values which unnecessarily increases the memory and computation load e.g.

among one lakh categories, maybe 70,000 categories have got no values, whereas only 30,000 categories have values. So, this kind of computation then takes a lot of memory and the entire computation process becomes slower. To overcome this issue a list structure can be used instead of a matrix form.

A sparse list reduces memory requirement by considering a list-based representation of attribute details which is proportional to the size of the seed and the number of attributes do not depend on categorization details i.e. each of the 30,000 demographic groups (with non-zero values) are based on certain attributes or variables and only those variables are listed and the number of individuals in those particular groups is listed which could be also called the weight of that particular group. This weightage could be again multiplied with an expansion factor like the proportion that had been estimated in IPF.

For example, in an ILUTE synthesizer, classical IPF is used and a real-valued expansion factor is attached directly on the list of attributes for each of those demographic groups.

(Refer Slide Time: 16:24)

Allocation Synthetic reconstruction

- ❑ IPF helps in estimating the joint distribution of the different variables which are eventually used to build the models.
- ❑ First the weights/proportions/values of each cell are adjusted to integers. This is called integer conversion.

Synthetic population of households:

- ❑ Select households from the reference sample in proportion to the estimated probabilities given in the multiway table obtained by the IPF or other technique.
- ❑ Selection with replacement is more common than selection without replacement.
- ❑ All variables essential for building the models are retained.

Number of household of each type can be determined by:

1. Total number of households X Probability in the multiway table (as per IPF)

Household selection from sample

- Monte Carlo (random) sampling from conditional probabilities (multiway table)
- Sequential procedure (Household heads → age, sex, marital status)
- Agents with higher probabilities appear more in the synthesized population

Synthetic reconstruction

IPF helps in estimating the joint distribution of the different variables, which are eventually used to build the weights or proportions or actual values of each cell of the cross-classification table. These values are adjusted to integers. This process is called integerization or integer conversion (explained in lecture 19) and this results in some problems because of rounding up.

At first, the weights or proportion or values for each of those cells i.e, for each of the demographic groups have to be determined which can be used to create a synthetic population of households or individuals also. For example, for household-level population reconstruction, households from the reference sample (i.e. disaggregate sample) should be chosen in proportion to the estimated probabilities given in the multi-way table obtained by IPF or other techniques. There are also other techniques for estimating this weight/proportion values and using this the probabilities can be estimated, and based on these probabilities, the samples can be selected from the disaggregate sample set.

Selection with replacement is more common than selection without replacement. i.e, from a sample list after selecting one sample, the next sample is selected with an assumption that in the next draw, there is an equal chance of choosing the previous sample again. But in many cases by using some measures it can be applied that in the next draws, a sample can have a lesser chance of getting selected from a particular group which could be done by attaching some extra weights to these particular probabilities.

All variables essential for building the models also have to be written i.e. in multiple cross-classification tables, single cross-classification tables can be combined where the probabilities can be estimated by doing joint probability estimations, or some of the variables in a table can be selected instead of all variables in the table.

So, the number of households of each type can be determined by the total number of households multiplied by the probability in the multiway table. Household selection from the sample is done by using Monte Carlo method using the conditional probabilities from the multiway tables and this process could be also done in a sequential manner. For example, if individuals are selected based on only age and sex. The conditional probability of a individual of particular sex given a particular age will be determined. If it is done for age, sex and marital status, the probability of marital status can be determined with the condition that age and sex group is of a certain type. So, using the sequential procedure, the probability of a particular group or probability of a particular sample to be determined.

One problem is that, calculating probabilities in sequence may result in varying results as per the sequence followed. For example, the result following the sequence, marital status, then age, and then sex differs from the result obtained if the sequence is sex followed by age and marital status. Usually, a sequence is followed based on the decision criteria in which that particular synthetic population is drawn for a particular model. For example, in the case of a household level mode choice model, probably at first the household head is required to determine followed by his employment status, location of office if employed, and finally, mode choice for a work trip. Hence, the sequence is determined based on the requirement. Agents with higher probabilities appear more in the synthesized population in Monte Carlo simulation.

(Refer Slide Time: 21:30)

SERIAL NO	WEIGHT	Household size	Household type	Other attributes
1	6	2	Family: married couple	...
2	9	3	Family: married couple	...
3	18	4	Family: married couple	...
4	18	1	Nonfamily: female living alone	...
5	6	1	Nonfamily: male living alone	...
6	6	2	Family: married couple	...
7	15	2	Family: female householder	...

SERIAL NO	Probability	Cumulative Probability
1	0.089	0.000
2	0.150	0.239
3	0.300	0.539
4	0.113	0.651
5	0.038	0.689
6	0.089	0.778
7	0.222	1.000

1. The conditional probability of an event B given that an event A has already occurred. This is written $P(B|A)$ or probability of B given A.

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

Where,

$P(A \text{ and } B)$ or $P(A \cap B)$ is the probability that both events A and B occur.

2. Cumulative probability is determined.

3. A random number is drawn(0 to 1).

**4. E.g., Random value = 0.705.
HH with S.NO = 5 is selected.**

5. HH is added to the target area and count tables are updated.

6. Individuals of selected households form the synthetic population.

7. Individual totals may not match.

For example, in the sparse list structure table (above figure) there are different household types, family and a married couple with various household sizes (2, or 3), non-family with female living alone, and non-family with men living alone etc.

The weightage of 6 means that there are 6 numbers of this type of household. For example, the weightages of the family with married couple is 6. So, if a particular sample is drawn from this particular table, then automatically the weight will be reduced from 6 to 5 and in this way, it can be recorded that one sample has been drawn of this particular type. This could be also used in probability calculations later on.

The Conditional probability of each type of group is estimated. The Conditional probability of an event can be estimated by using the following formula. For example, the conditional probability of an event B given that an event A has already occurred can be estimated by using the following formula.

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

Where,

$P(A \text{ and } B)$ or $P(A \cap B)$ is the probability that both events A and B occur.

For example, the conditional probability of sample with household size 2 given that they are married couple and family is 0.089. It can be said as the conditional probability of the married couple given that they are family with household size 2.

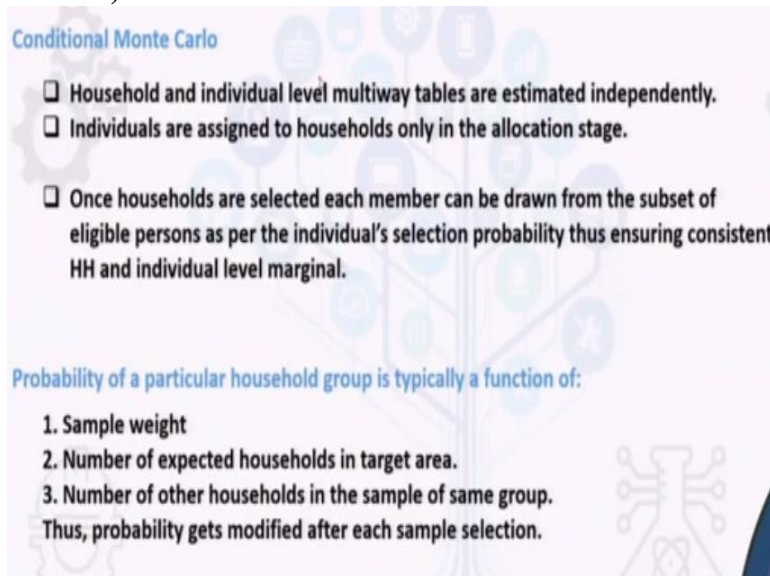
Alternatively, the weight can be used for calculation of the probabilities which can be used if there are only 2 variables. But if there were more than 2 variables and for each different weightages are there, then using weightages, the control totals of each of these groups, and the intersection of those 2 groups, conditional probabilities can be also determined.

After the determination of the conditional probabilities of each of the samples, the listing will be done. Then the cumulative probability (in the above figure) will be determined by adding them sequentially. For example, by adding the probability of first and second groups the second cumulative probability will be found (e.g. $0.089+0.150 = 0.239$) with which the probability of the third group will be determined (e.g. $0.239+ 0.300 = 0.539$).

Then a random number is drawn from 0 to 1. For example, if a random number of 0.705 is drawn it lies in the 5th group because the next group starts from 0.778 i.e. household number 5 is selected from the sample. This is the process of drawing a particular household from the sample and then this will be put it in their actual target area and will be continued till the total number of population for the targeted area is fulfilled, provided that the individual variables (i.e. individual characteristics of the household) match with their control totals in the target area.

If households are selected to form the synthetic population sometimes individual totals may not match i.e. as the household is selected, and based on each household the individuals are determined, the individual totals may not match with the areas control totals for individuals, but the household totals will match. However, there are other procedures where both of them could be also matched by using different weightages or using both of them together and then giving weightages to both. So this is the simplest way to determine a synthetic population for a particular zone.

(Refer Slide Time: 27:26)



Conditional Monte Carlo

- ❑ Household and individual level multiway tables are estimated independently.
- ❑ Individuals are assigned to households only in the allocation stage.
- ❑ Once households are selected each member can be drawn from the subset of eligible persons as per the individual's selection probability thus ensuring consistent HH and individual level marginal.

Probability of a particular household group is typically a function of:

1. Sample weight
2. Number of expected households in target area.
3. Number of other households in the sample of same group.

Thus, probability gets modified after each sample selection.

Conditional Monte Carlo

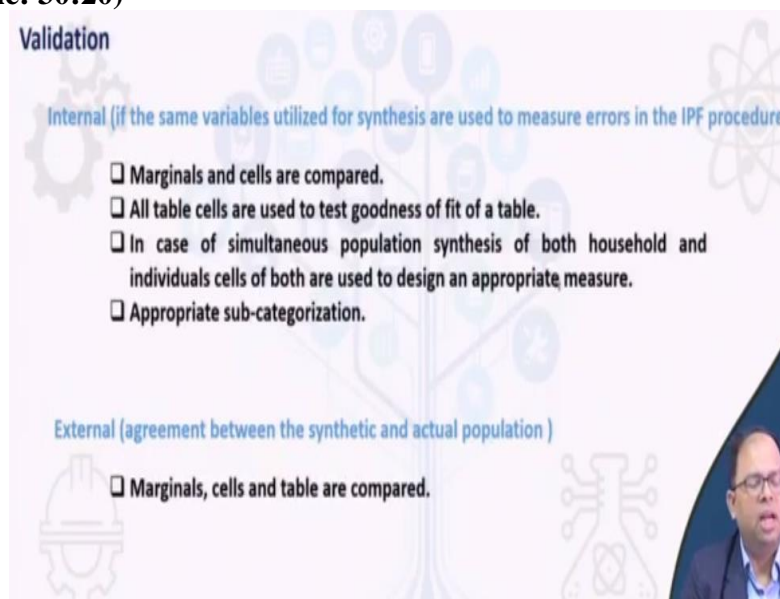
The probability of a particular household group can be determined by using conditional probabilities or using a function which includes the sample rate or the number of expected households in the target area i.e. the number of households in that target area may play a role in the total number of sample drawn.

Similarly, the number of other households in the sample of the same group (i.e. the weightage of a particular sample of particular demographic type) or a particular group in a particular sample (i.e. the disaggregate sample) may play a role. So, by using these concepts, a function can be developed to determine the probability of a sample of getting selected which is used in more complex software or procedures.

After every selection, the probability will get modified for the next sample selection i.e. the probability need to be recomputed after every selection that will also change the cumulative probability, and then another sample will be drawn.

In conditional Monte Carlo, households and individual level multiway tables are estimated independently i.e. both the cross-classification tables for households and individuals are done separately and then individuals are assigned to households only in the allocation stage i.e. during the selection of the households for a particular target area, the types individuals selected is also analysed followed by matching the totals for both and then the particular household is selected in the allocation stage. Once households are selected, each member can be drawn from the subset of eligible persons as per the individual selection probability ensuring consistent household and individual level marginal.

(Refer Slide Time: 30:20)



Validation

Internal (if the same variables utilized for synthesis are used to measure errors in the IPF procedure)

- Marginals and cells are compared.
- All table cells are used to test goodness of fit of a table.
- In case of simultaneous population synthesis of both household and individuals cells of both are used to design an appropriate measure.
- Appropriate sub-categorization.

External (agreement between the synthetic and actual population)

- Marginals, cells and table are compared.

Validation

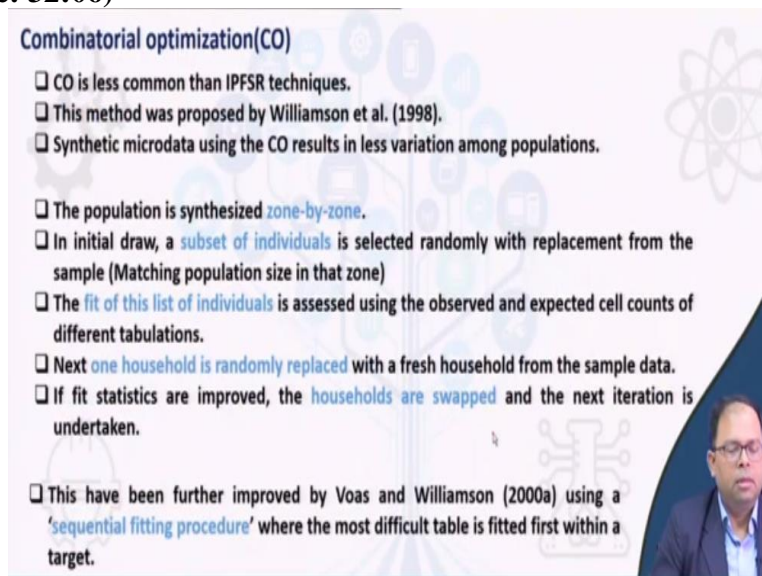
After completion of the IPF procedure, there is a need to validate the results. The validations could be done internally or externally. Internal validation means if the same variables are utilized i.e. the variables used for synthesizing the population are also used to measure the errors in the IPF procedure i.e. marginals and cells are compared. So, for the target population and final

results, all table cells are also used i.e. the goodness of fit test by using all the cells (estimated findings in each cell and actual figures) can be done.

In the case of simultaneous population synthesis, households, and individual cells, both could be used together to design an appropriate measure and new goodness of fit measure can be used where both things would be included.

Appropriate subcategorization is important to get the best possible synthetic population. External validation means the agreement between the synthetic population and the actual population from the survey. Hence, the actual population data must be available. Here again, marginal cells and tables are compared, but in case of the absence of actual population data to validate, validation procedure is limited to internal validation only.

(Refer Slide Time: 32:06)



Combinatorial optimization(CO)

- ❑ CO is less common than IPFSR techniques.
- ❑ This method was proposed by Williamson et al. (1998).
- ❑ Synthetic microdata using the CO results in less variation among populations.
- ❑ The population is synthesized zone-by-zone.
- ❑ In initial draw, a subset of individuals is selected randomly with replacement from the sample (Matching population size in that zone)
- ❑ The fit of this list of individuals is assessed using the observed and expected cell counts of different tabulations.
- ❑ Next one household is randomly replaced with a fresh household from the sample data.
- ❑ If fit statistics are improved, the households are swapped and the next iteration is undertaken.
- ❑ This have been further improved by Voas and Williamson (2000a) using a 'sequential fitting procedure' where the most difficult table is fitted first within a target.

Combinatorial Optimization (CO)

The combinatorial optimization, proposed by Williamsons in 1998 is less common than the IPFSR technique. It has been found from many literatures that synthetic microdata using CO results in less variation among the population. So, this is a more robust procedure compared to the IPFSR technique, but it is more computationally intensive and the population has to be synthesized zone by zone i.e. it has to be done for each zone separately.

At first, a subset of Individuals is selected randomly with replacement from the sample i.e. a subset of individuals is selected randomly from the disaggregated sample. The total number of samples drawn must match the population size for the zone for which it has been selected.

Then, the fit of this list of individuals is assessed using the observed and expected cell counts of different tabulations i.e. the number of males, females, individuals in each age group, will be measured and compared with the marginals for that particular zone. This is done by estimating the fit statistics which shows how much different is the selected sample from that expected values.

Then one household from the selected sample is replaced randomly with a fresh household from the sample data and it will be continued till further replacement does not improve the fit statistics anymore. If the last replacement improves the fit statistics, it will be retained, otherwise, the new household will be returned and another household from the sample will be selected. So, if fits statistics are improved then, households are swapped and the next iteration is undertaken.

It has been further improved by Voas and Williamson using a sequential fitting procedure, where the most difficult table is fitted first within a target i.e. among several classification tables the most difficult ones are first fitted and then the easier ones are fitted.

(Refer Slide Time: 35:22)

Step 1: Sample survey and target area marginal/constraints

Survey micro data				Known small area constraints			
Household	Characteristics			1. Household size (persons per household)		2. Age of occupants	
	size	adults	children	Household size	Frequency	Type of persons	Frequency
a	2	2	0	1	1	adult	3
b	2	1	1	2	0	child	2
c	4	2	2	3	0		
d	1	1	0	4	1		
e	3	2	1	5+	0		
Total				2		2	

Step 2: Households are selected (as per constraint i.e., 2 here) randomly from sample(a,e) as initial microdata set for small area

Step 3: Households are tabulated and (absolute) difference is calculated

Household size	Estimated Frequency(i)	Observed frequency(ii)	Absolute difference(i)-(ii)	Age	Estimated Frequency(i)	Observed frequency(ii)	Absolute difference(i)-(ii)
1	0	1	1	adult	4	3	1
2	1	0	1	child	1	2	1
3	1	0	1	Sub total:		2	
4	0	1	1	Total absolute difference		(4+2)=6	
5+	0	0	0				
Sub total:				4			

(Source: 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))

In this example (in the above figure), in step 1, suppose A B, C, D, E are household samples with characteristics like household size, number of adults and number of children in that particular household . Sample number A is of size 2 with 2 adults and there are no children. The sample number B has a household size of 2 with 1 adult and 1 child. This is the survey micro data or disaggregate data.

Next, are the control totals for that particular target area. Targets are constraints which can be called the marginal targets constraints e.g. for household size 1 there is one household and for households size 4 there is also one household and the total target area has got only 2 households. Similarly, this target area has got 3 adults and 2 children.

There are no houses with household sizes 2, 3, and 5 plus. So, the total number of houses is 2 and the total population is 5. Then households are selected randomly from the sample as initial micro data set for the sample. So, as per the constraint, there are only 2 houses i.e. 2 households will be selected from sample data.

Two households have been selected randomly from the sample and suppose a and e is selected. Then these households are tabulated (step 3 in the above figure) and the absolute difference is calculated and this absolute difference is the fit statistics e.g. the estimated frequency for household size 2 and household size 3 will be one. However, the actual frequency for household size is 0 and for household size 1 and 4, is one. So, absolute differences is 4 (1 + 1 + 1 + 1) i.e. there are 4 cases of mismatch.

While looking at individuals, the selection of a and e results in the estimated frequency of adults as 4 and the estimated frequency of child as one whereas the observed frequency is 3 and 2. So, here again, there is an absolute difference of 1 and 1. So, the total absolute difference is 2, and adding up this to absolute differences results in a total absolute difference of 6.

(Refer Slide Time: 38:56)

Step 4: One of the pre-selected households is randomly selected and replaced with another randomly selected household from the sample only if (absolute) difference is reduced (d replaces a)

Household size	Estimated Frequency(i)	Observed frequency(ii)	Absolute difference (i)-(ii)	Age	Estimated Frequency(i)	Observed frequency(ii)	Absolute difference (i)-(ii)
1	1	1	0	adult	3	3	0
2	0	0	0	child	1	2	1
3	1	0	1				
4	0	1	1				
5+	0	0	0				
	Subtotal:		2				
					Total absolute difference		(2+1)=3

Step 5: Step 4 is repeated till no further reduction is possible (d and c: final)

Household size	Estimated Frequency(i)	Observed frequency(ii)	Absolute difference (i)-(ii)	Age	Estimated Frequency(i)	Observed frequency(ii)	Absolute difference (i)-(ii)
1	1	1	0	adult	3	3	0
2	0	0	0	child	2	2	0
3	0	0	0				
4	1	1	0				
5+	0	0	0				
	Subtotal:		0				
					Total absolute difference		(0+0)=0

(Source: 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))

In step 4, one of the pre-selected households is randomly selected and replaced with another randomly selected household from the sample only if the absolute difference is reduced. For example, a is replaced with d and now the total absolute difference comes down to 2 for the households, and for the population, the absolute difference comes down to 1. So, the total absolute difference is 3. It continues until there is no further reduction of this absolute difference. Hence, d and c are the final selections for this particular zone.

(Refer Slide Time: 39:46)

Relation matrix

ALBATROSS

- A relation matrix is used to estimate household composition that matches individual-level constraints.

Contingency table (two rows and two columns)
 e.g., single & male-female households
 (couples, independent male, independent female, total male, total female)
 Additional row + column in the contingency table for including children.
 Disaggregating (sub-matrices) by age and work status.

- Matrices are seeded using the sample.
- IPF is undertaken on these relation matrices and the distributions obtained are used as marginal for conducting IPF of household variables.

Relation Matrix

In the Albatross model (discussed in the previous lecture) a relationship matrix is created to estimate household composition that matches individual level constraint. This is a new method.

Unlike the IPF method, the relationship matrix is first completely determined e.g. suppose a contingency table of 2 / 2 columns is created, where these 2 / 2 columns include single and male, female households. In these contingency tables, the data related to couples, independent male, independent female, total male, total female are available. Now, another row can be added to this contingency table of 2 rows and 2 columns to add children for each of these categories by both the row and the column side, and then it can be further disaggregated this by age and work status of the head of the household. So, the size of the contingency table can be kept on increasing by desegregating data.

In this kind of a contingency table, both the population as well as the household type is combined i.e. in rows and columns different types of data for individuals as well as families are present. The sample data could be used as seed for these matrices. Then, IPF is undertaken on these relation matrices and the distributions obtained from this IPF are used as marginal for conducting IPF of household variables i.e. as the seed for households as well as individuals in those households are included when households are evaluated the control totals of individuals will also be evaluated automatically.

(Refer Slide Time: 42:16)

Iterative Proportional Updating (IPU) PopGen Software

- ❑ The IPU algorithm simultaneously controls both agent (individual-level) and agent group (household) level attributes by iteratively adjusting and reallocating weights.
- ❑ The data structure followed is a tabular list of agent groups similar to the sparse list variation of IPF.

Markov chain Monte Carlo (MCMC) simulation-based approach

- ❑ In this approach, the focus is more on to meet the partial joint distribution of population characteristics rather than matching the marginal sums (as in IPF).
- ❑ Joint distributions of population characteristics are organized in a Markov-Chain.
- ❑ In Markov Chain Monte Carlo sampling method, samples are drawn such that the next sample is dependent on the existing sample.
- ❑ Gibbs Sampling and Metropolis-Hastings algorithm are popular approaches to Markov Chain Monte Carlo sampling.

Iterative proportional updating (IPU)

Iterative proportional updating (IPU) is a variant of the iterative proportional fitting, which is used in the PopGen software, where the IPU algorithm simultaneously controls both agent at the

individual level and agent group at the household level by iteratively adjusting and relocating the weights.

The weights can be changed based on the selection of a particular household or a particular individual from that particular household. It can be done iteratively and the population group for the synthetic population for a particular zone can be determined. The data structure followed is a tabular list of agent groups similar to the sparse list variation of IPU. In short, there are agent groups (e.g. households), a tabular list, and individuals for each of these households. Weightage can be given based on both the values.

Markov chain Monte Carlo simulation

In this approach the focus is more on to meet the partial joint distribution of population characteristics rather than matching the marginal sums like in IPF i.e. how individual groups or individual demographic groups are distributed and the weightages of each individual demographic groups.

A simulation is done using those distributions and their dependencies which is not exactly a simple random Monte Carlo simulation. Here, joint distributions of population characteristics are considered and these are organized as Markov chains. In the Markov chain Monte Carlo sampling method, samples are drawn such that the next sample is dependent on the existing sample i.e. conditional probabilities are considered.

Hence, the selection of a particular sample and a particular attribute of a particular sample influences selection chances. So, it creates a selection chain and that is why it is called a Markov chain. There are two algorithms one is called a Gibbs sampling and Metropolis-Hastings algorithm, which are popular approaches to do Markov chain Monte Carlo sampling. This is another upcoming method that is also used in many new approaches to create synthetic population.

(Refer Slide Time: 45:24)

REFERENCES

- Alaska Department of Labor and Workforce Development. 2008. Iterative proportional fitting for a three-dimensional table (<https://edyhsgr.github.io/eddieh/IPFDescription/AKDOLWDIPFTHREED.pdf>)
- Iterative Proportional Fitting Procedure (IPFP)(<http://www.real-statistics.com/matrices-and-iterative-procedures/iterative-proportional-fitting-procedure-ipfp/>)
- Beckman, R.J., Baggerly, K.A., McKay, M.D. Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice 30, 415-429.
- 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 p of firms (2009). Geographical Analysis, 41 (2), pp. 181-203.
- Guo, J. Y., & Bhat, C. R. (2007). Population Synthesis for Microsimulating Travel Behavior. Transportation Research Record, 2014(1), 92-101.

(Refer Slide Time: 45:29)

CONCLUSION

In India, since micro sample data surveys are not conducted by the government, we need to collect extensive sample datasets for each study and study area which is a major limitation in developing disaggregate models or conducting microsimulation studies.

While several software has been developed for population synthesis, most are custom built for specific purpose and geographical context. None exists for India.

IPF could be used for developing cross-classification tables(for agent and agent groups) for an urban area which could be used to develop disaggregate models.

IPFSR and CO techniques are popular and easy methods of population synthesis which could be easily adopted.

Conclusion

In India, as micro-sample data surveys are not conducted by the government, there is a need to collect extensive sample datasets for each study area, which is a major limitation in developing disaggregate models or conducting microsimulation studies.

Then, while several software has been developed for population synthesis, most are custom built for a specific purpose and geographical context and none exists for India. So, there is a need to develop this kind of software in the Indian context so that micro simulation procedures, detail land-use transportation models or any disaggregate models can be performed. Then, IPF could be

used for developing cross-classification tables for agents and agent groups for an urban area which could be used to develop discrete choice models.

Based on the prediction model requirement other variables can also be added into the IPF procedure to have more categories or more cross classifications.

IPFSR and CO techniques are popular and easy methods of population synthesis which could be easily adapted.

Thank you.