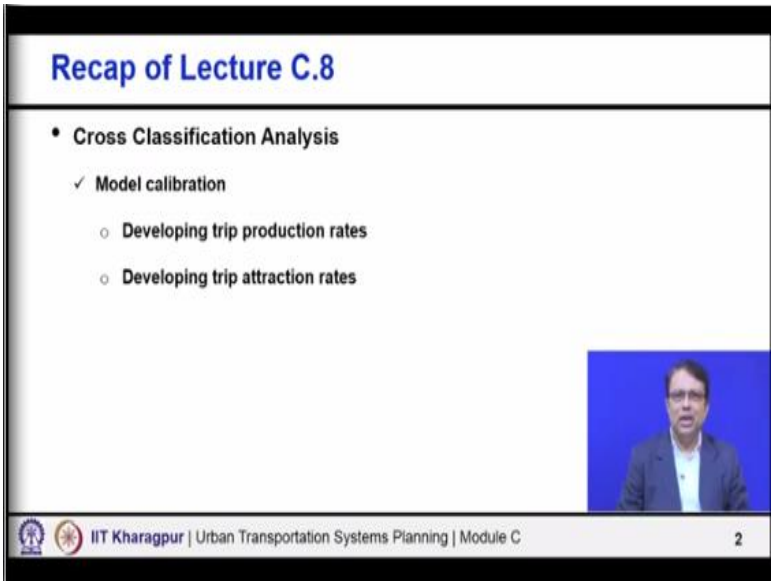
Urban Transportation Systems Planning
Prof. Bhargab Maitra
Department of Civil Engineering
Indian Institute of Technology-Kharagpur

Lecture-19
Cross Classification Analysis:
Model Application, Advantages and Disadvantages

Welcome to module C, lecture 9.

(Refer Slide Time: 00:22)

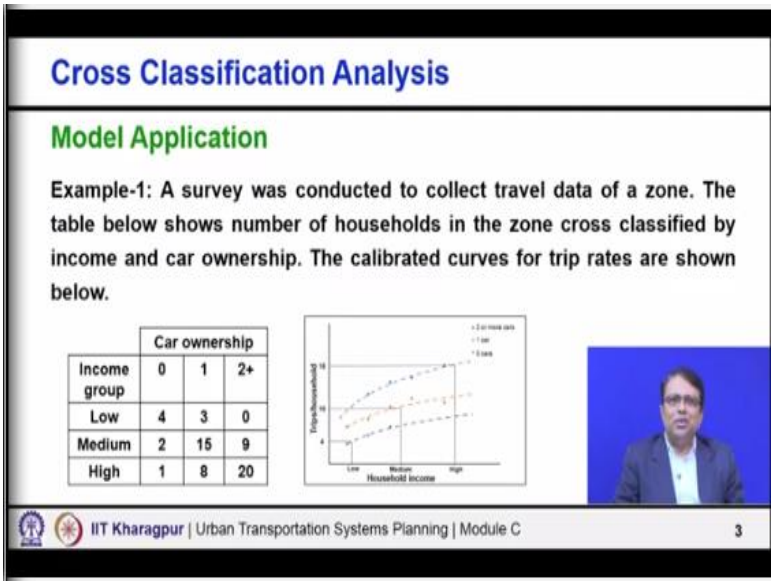


In lecture 7, we introduced to you about the cross classification analysis, what we do, what are the fundamental assumptions. Then in lecture 7 also we covered what should be the structure, trip generation, production and attraction separately what would be the model structure. Then in lecture 8, we discussed about the model calibration part that means once we have developed the structure as per the structures how to develop the trip rates?

And further classify those trips for different trip purposes or different categories of trip etc. So, that discussion was made particularly the model calibration part in lecture 8. And we explained to you how to develop the trip production rates and trip attraction rate based on the model structure. So, we define the model structure then we develop the rate and why we did all these things, ultimately purpose what is our ultimate purpose?

That ultimate purpose is to apply those for future scenario. So, in this class, I am going to discuss about the application part, how to apply these trip rates? Whatever we have developed using the cross classification techniques for forecasting trip productions and attractions for a given zone or studied here.

(Refer Slide Time: 02:10)



So, let us come to the model application part. The first example I have taken is like this, a survey was conducted to collect data of a zone. For a particular zone the table below shows the number of households in the zone cross classified by income and car ownership, the calibrated curves for trips rates are shown also in this slide. So, these calibrated curves are very similar to as you have seen during the calibration.

So, based on calibration, we considered 3 households let us say 3 or 5 household income and 3 car ownership level. But let us say in this example, we are considering 3 household incomes. So, suppose we know that low, medium and high, so it is not directly that same values we are borrowing not a continuation of the previous example. But an independent example and we made this example simple one**.**

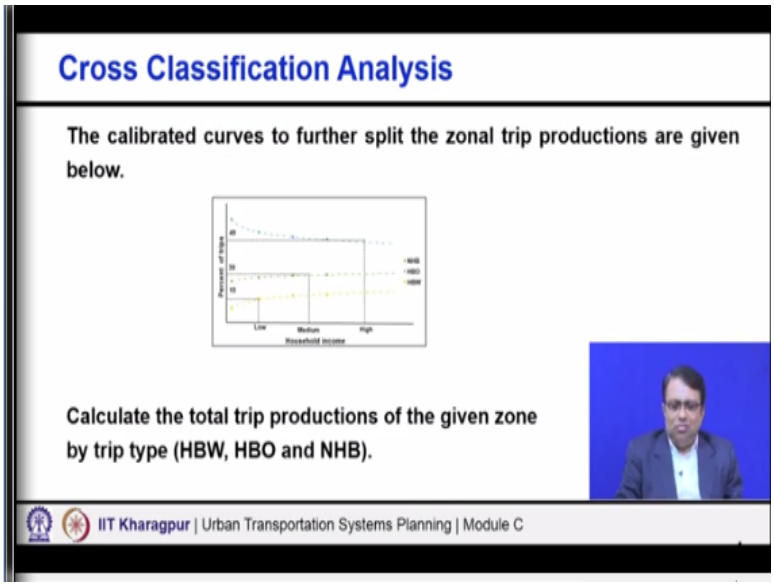
Because our again objective here is to explain you how you can apply it. So, in that example we considered 5 categories based on the a trip household income and 3 based on the car ownership. But here in this example, we are using it only 3 income categories say low income group,

medium income group and high income group. Let us take that the rates have been developed as per this groups.

So, the vertical lines you can see clearly, this represents the low income group, this represents the medium income group and this represents the high income group. Then for different car ownership, we got the data, now tag how many households are there in each category that is given. And for each category that means 3 income groups and considering 3 car ownership levels.

We also know the trip rates which are explained here, each car represent one particular car ownership category. And then low, medium, high x axis represents the income group and y axis value represents trip per household. So, we know that trip rates, we know the number of households, those who are there.

(Refer Slide Time: 05:23)

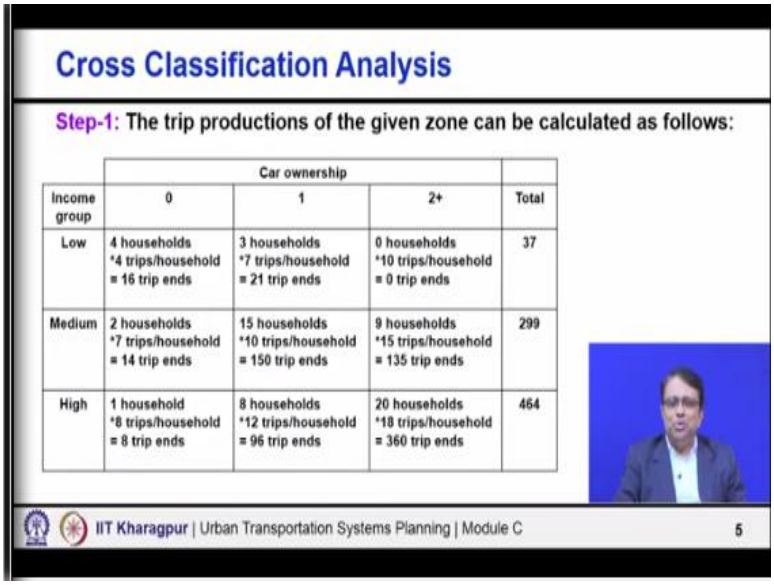


Then we also have another curve which is given to get the share of trips considering different trip types, say home based or how home based work, home based other and non home based, these are the 3 categories we have considered in this example. So, for each category again here also x axis represents low, medium and high, 3 income groups. These vertical lines are indicating those income groups and then there are 3 curves, one is for home based work, red one.

And then the other 2 are home based other and non home based trip, so y axis represent percentage of trips. So, the first one the trip rates for each category, you can get the total number of trips in that zone and then this curve can help us to get the percentage of total trips under home based work, home based other and non home based trip. So, what do you want here? We want to see how we can calculate the total productions of the given zone by trip type.

Trip type means, how many are home based work, how many are home based other and how many are non home based trip? So, let us look at the solution part.

(Refer Slide Time: 07:07)



When we look at the solution, we have created again those cells, so we had 9 cells, 3 income group and 3 car ownership. Each case the 2 inputs are required for me, one is how many households are there in that category, and number 2, what is the trip rate for that category of households? So, the number of household, if you go to the earlier slide, we know that 4, 3, 0, 2, 15, 9, 1, 8, 20 these are the respective number of households in zone 1 under different categories.

So, the number of households (()) (07:53), how we get the trip rate? Trip rate we know, this is each category for low income and 3 curves, middle income 3 curves for 3 car ownership group, high, again 3 curves, 3 points. So, here we get 3 points, for medium income group we get 3 points, for low income group we get 3 points. So, for each of these categories, low, medium, high and based on 3 levels of car ownership, we know the trip rates from this table.

So, using these 2, we go back here, what we are doing here? So, let us say first category, we know there are 4 households. And as per those trip rate curves low income with no car ownership make 4 trips per household, so total how many 4 into 4. Similarly shear low income and one car we have got 3 households in that category. And from the curves, we find that the applicable trip rate is 7 trips per household.

So, how many trips we had? 3 into 7, 21. Exactly following the same procedure for low income, but 2 + household we have got 0 that is also unrealistic. One way we are saying low income, but then 2 and more cars, that is not possible. So, we have zero households, so zero number of trips. So, how many total trips we get from low income group 16 + 21, so total is 37. Medium income group again following the same thing, the respective number of households are 2, 15 and 9 and respective trip rates are 7, 10 and 15.

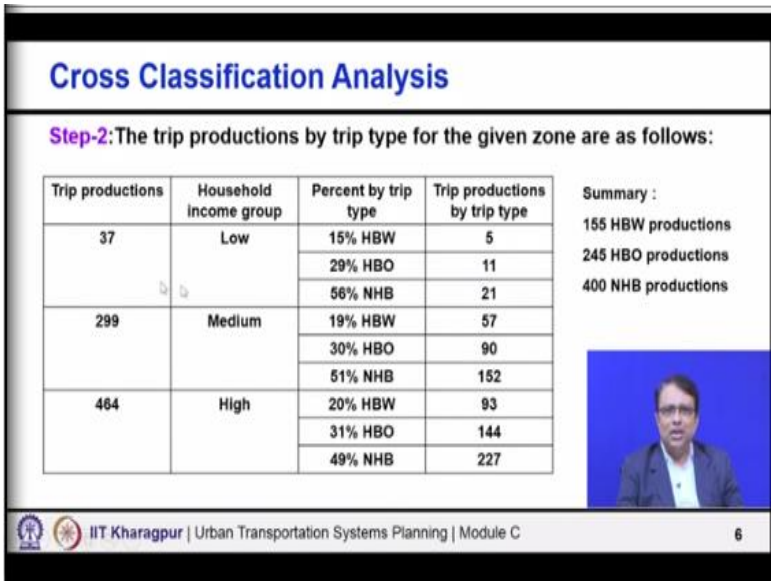
So, the respective number of total trips are 14, 150 and 135, you add all the 3 together you get 299 trips, that will be produced by the medium income households. From high income household, the respect household numbers are 1, 8 and 20 under 0 car ownership, one car and having 2 or more cars and the respective trip rates are 8, 12 and 18, so you get 8, 96 and 360 trips respectively total 464.

So, I know given this trip rates or established trip rates in the future if I have so many households under each category. Then I can estimate the low income household will make total 37 trips, medium income household will make 299 trips and high income households will make 464 trips. So, you can add these 3 numbers to get the total number of trips that are likely to be produced from that zone.

Now we also have another input given to us for each income group whatever total trips are produced, how much percentage will be home based work, home based other and non home based trip. Now we take this percentage, so whatever total number of trips have been produced 37. So, for low income group, whatever total trips are produced, we know the percentage of total trips which will represent home based work, home based other and non home based trip.

So, this 37, on 37 we can apply those percentages to calculate the split of this 37 under 3 categories. Same thing I can do for middle income households and for high income households.

(Refer Slide Time: 12:19)



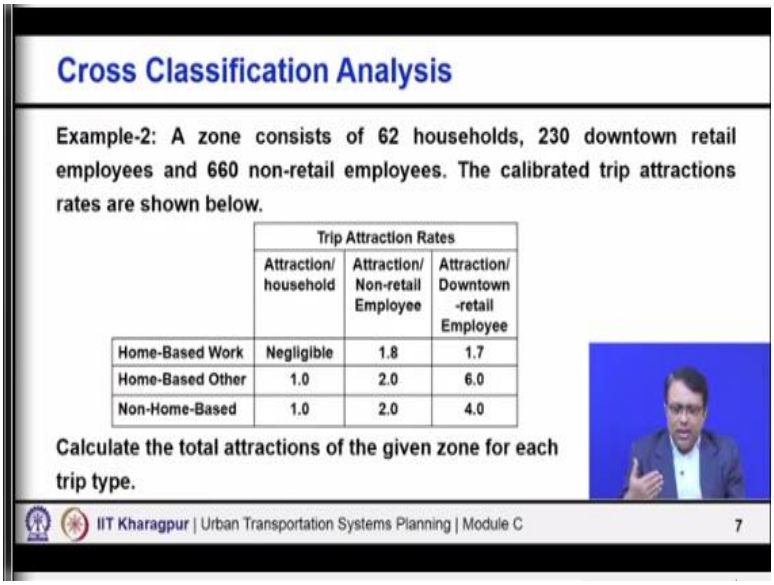
So, the next slide we have said that, so we know that the low income household will produce 37 trips. But from the second curve, second curves means which one I will show you again, again this curve. So, we know that for low income household 15% will be home based work, 29% will be home based other and 56% will be non home based trip. So, that curves gives us these values, those curves will give us these values.

So, accordingly 37% 37 trips, 15% is about 5 trips, then 29% about 11 trips, 56% about 21 trips. So, we can tell you how the trips are going to be by type of trips. Exactly the same thing we apply now for 299 only this percentage values will be as per the medium income group household, and the next this percentage will be as per the high income households, so you get the number.

So, from this one, you will be able to say in the future how many trips will be produced in total from zone 1, and how much low income household will produce, how much medium income households will produce, how much high income household will produce? And then each case what will be the share of home based work, home based other and non home based trip and then

all total from the zone how many home based work trip will be produced, home based other trips will be produced and how many non home based trips will be produced? So, that is all those calculations we can get, that is one application.

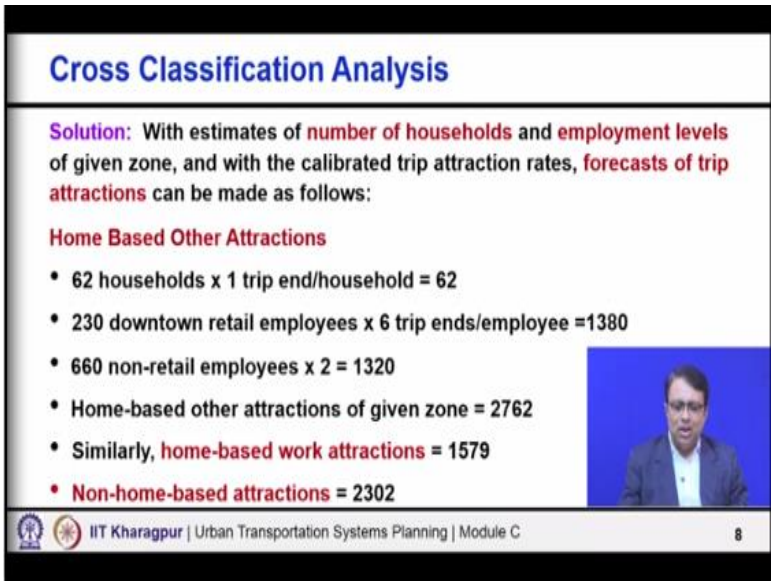
(Refer Slide Time: 14:11)



Let me take another example which shows the application for the trip attraction. In this case, we are saying zone consists of 62 households, 230 downtown retail employees, and 660 non retail employees. So, 62 households, 230 downtown retail employees and 660 non retail employees, the calibrated trip attraction rates are given below. So, you can see that the trip attraction rates are given for home based work home based other and non home based, and then 3 types of employment or attractions, what are the rates?

One cell it is negligible, so you can assume it to be 0, the other cases that rates are given. Now what do we want? Calculate the total attractions of the given zone for each trip type. Trip type means, how many home based work, how many home based other and how many non home based? That is what we wanted. And I know that there are 62 households 230 downtown employees and 660 non retail employees, those may be the forecast for a horizon here.

(Refer Slide Time: 15:44)



So, what we do? with estimated number of households and employment levels of given zone and with the calibrated trip rates exactly as we did for the trip productions same way we try to forecast the trip attractions, how you do that? Let us say first home based other attractions if you take, we have 62 households and we have one trip per household. Now home based other, home based other means.

Home based other will have 3 component attraction per household, then attraction per non retail employee and attraction for downtown retail. This attraction per household also we say these are not household retail for it is not their residence of the trip maker other than the resident of the trip maker. So, this 1, 2 and 6 these rates are known, so we can get 1, 2 and 6 and then 62 we have 660 non retail employees, and 230 downtown retail employees.

So, accordingly you can get 62 into 162 trips, 230 into 6 = 1380 trips, and 660 non retail employees multiplied by 2, so 1320. So, that total home based other will be how much? Home based other considering these 3 components with trip rates 1, 2 and 6, that is what we did. And total then will be how much? Total will come 2762. So, exactly following very similar calculations using these numbers 62, 230 and 660 using this numbers but using the appropriate rates, we can find out also home based work and non home based trip that is what we presented here.

So, non home based work attraction comes out to be like this, the number is not important, but how we are calculating, how we are applying that is what is of interest to us. And then also same way we calculate non home based attractions. So, each case we use the household 62 household, 230 downtown retail employees, 660 non retail employees. But this 1, 6 and 2 in this case, in other case we use different rates, for non home based we use 1, 2 and 4 rates respectively and home based work first case 0, second 1.8 and then 1.7.

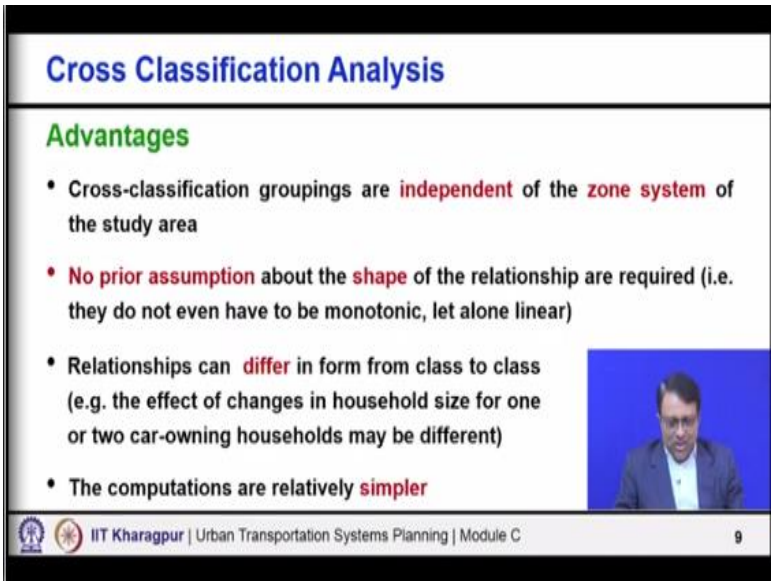
So, these rates will change as per the type of trip. So, you can actually tell, then in the future if these are my number of household, these are that likely number of downtown retail employee and this will be their number of non retail employees, then there will be so many home based work, home based other and non home based trips, that are likely to get attracted to this zone. The same way or following the similar procedure, I have shown it for one zone.

So if a study area now has maybe 50 zones or say 20 zones, then for all 20 or 50 zones, you can calculate how many trips will get attracted and then how many trips will get produced in the first example? And then using this example in each of those zones how many trips are likely to be attracted. And in this case we have classified into home based work, home based other and non home based trip.

So, in each case you know how many are going to be produced in each zone, how many are going to be attracted in each zone. So, that means my the productions attractions by category wise, trip category wise or trip classification wise, I know the way I want to classify. Now as you said here home based, non home base, home based other we can use it purpose wise, we can use it mode wise, so many ways you could express these trips**.**

Or if nothing is there at least the total number of trip productions and total number of trips attractions, that for all the zones we could say that, that is what it is.

(Refer Slide Time: 20:51)



Now I would like to discuss about the advantages and the disadvantages associated with the cross classification analysis. You have now you know 2 types of approaches very clearly, one is the regression based model or the regression model I can say, that is one way of doing the trip productions and attractions or the trip generation modeling. The other approach you have seen how we do with what assumptions, and then basically the cross classification analysis or category analysis.

Now both things are known, so, it is the right time for us to look at these 2 approaches and try to understand the merits and demerits. Since we now discussed only the cross classification, so we will try to highlight the advantages and disadvantages of cross classification technique. But obviously the comparison is being made easily you can understand and compare it with the other method.

So, let us see the first one is advantages, the cross classifications grouping are independent of the zone system of the study area. This was a major issue when we talked about the regression based model. If you remember my discussion, when we talked about the zonal based regression, we said that if you are using zonal total, then there are so many implications of using zonal total as a variable or zonal totals as variables.

And then reducing the zone size helps us to the sorry using the multiplier 1 by H i helps us to or using the rate 1 by H i means number of household in each zone, we divided it the total by that number of households in each zone, so you get the rate. If we are using the rate then we are making the model almost independent of the zone size that influence come out.

On the other hand, we said that if we reduce the zone then my inter-zonal variability increases, my intra-zonal variability reduces. So, these are advantageous but then we have higher resources are required. For model development, more expensive model data collection, model calibration, application, so those are their implications. So, always you say for the regression based model, the zone size and zone system at the back end somewhere it is influencing always.

But this cross classification groupings are independent of zone system in the study area, we do not bring zone system into consideration unless we are going for only thing in the last stage where we have want to apply. There we know that how many zone wise, how many households and what category or if it is extraction, then what kind of land use and the character and then quantity, those are coming for the forecast purpose.

But when we are calibrating our model is independent of the zone model. And what we are borrowing are basically the rates what we have established, which are absolutely without any influence of the zone. Second advantage, no prior assumption about the shape of the relationship are required, that means they do not even have to be monotonic let alone linear, what does it mean?

We assume the relationship whenever you want to develop regression model, first you assume that the linear relationship is linear, why is linearly related to X, X 1 or X 1 and y linearly related, X 2 and Y linearly related, that is the first assumption. So, in my model specification itself, I am assuming something and then with that assumption I am calibrating the model and to see how based my data fits to the model, my model fits to the data.

And then based on the R square and T value and all logical sign constant, we are checking that how good the variation in the data could be explained by the model. But we are making an prior

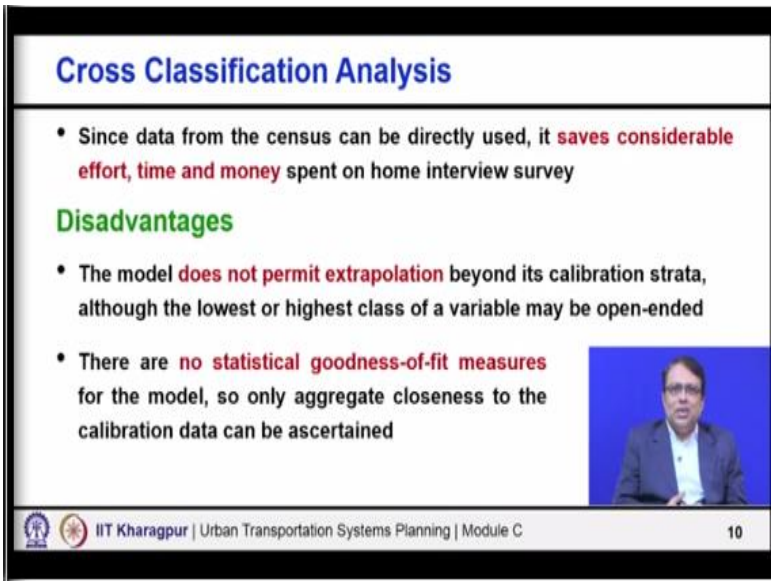
assumption. In this case we do not need to make any prior assumptions, let it come it may come linear, it may come nonlinear. Say for example, if 0 to 1 car and 1 to 2 car the effect need not be even linear.

So that kind of great advantage we can get from use of cross classification analysis. Third, relationships can differ in form from class to class, that means the effect of change in say household size for 1 or 2 car owning households maybe different. The curves whatever you plot, 1 car, 2 car, 3 car, they are not just parallel lines. So, if I say that for low income household you know another attribute changes, a level of another attribute changes.

Medium income household the same attribute level changes, the impact maybe different, the trip rates may differ. That is the overall flexibility, the rigidity is very less, in this sense there are problems as well which we will discuss. But in this context these are quite useful that prior assumption, no prior assumption, even the relationships can differ in from class to class. Third, the computations are relatively simple, you do not have to really go for any mathematical modeling or nothing is required you simply get.

So, many number of households you need only them to be statistically representative, so many households making so many trips. So, the total trips divided by total number of households considered under the category, get the rate, same thing is true in the context of trip attraction rates. So, the methodology is very simpler.

(Refer Slide Time: 28:28)



Then since the data from the census can be directly used, it saves considerable effort, time and money on home interview survey. Many cases, you get the data for the let us say the number of households, what is the income stratification, many of this data directly we can get from the secondary sources, census, economic census, so they can be used advantageously.

But not that all or everything is only good, all are good and there is no disadvantage that cannot happen. For every method there are advantages, there a disadvantage and that is why an understanding helps us to decide given a context, you were doing a one work, maybe look at the what is really the intention or the main objective of the work, how much fund is located, how much time is given.

And then decide which is the most appropriate method maybe in one project you will use regression based model, in another project you may come out with a decision that ok no, I want to use cross classification or category analysis. So, it has to be context dependent, because the method advantages and disadvantages are known to you. So, you are comparing the advantages and disadvantages and then trying to make a call for a given context, what is my best decision.

Everywhere, wherever there are alternatives for practically in every context in this world, any planning, any engineering, any decision making you do but alternatives are available. So, one has to make a decision, what is the best one? Let us quickly see on the disadvantage, the model does

not permit extrapolation beyond the classic calibrations strata, if you have considered 4 categories, that is all, you cannot create a fifth category later on.

So, whatever grouping you have done when you have calibrated, in future you have to use exactly the same grouping. Say for example, when I developed the work, then I think 80,000 income was considerably high and there were very further classification beyond 80,000 was not applicable. But maybe after some years, when you were applying this model you may feel, there are many people up to 82, 000, 100,000.

And maybe I want to consider them as a separate group, you cannot do it, because you have not calibrated. So, whatever we have calibrated exactly the same grouping you have to maintain, that is one part. Second, there is no statistical goodness of fit measures for the model maybe regression model, there were many assumptions, the assumptions were that variable I mean relationship is linear and so many other things we had to assume.

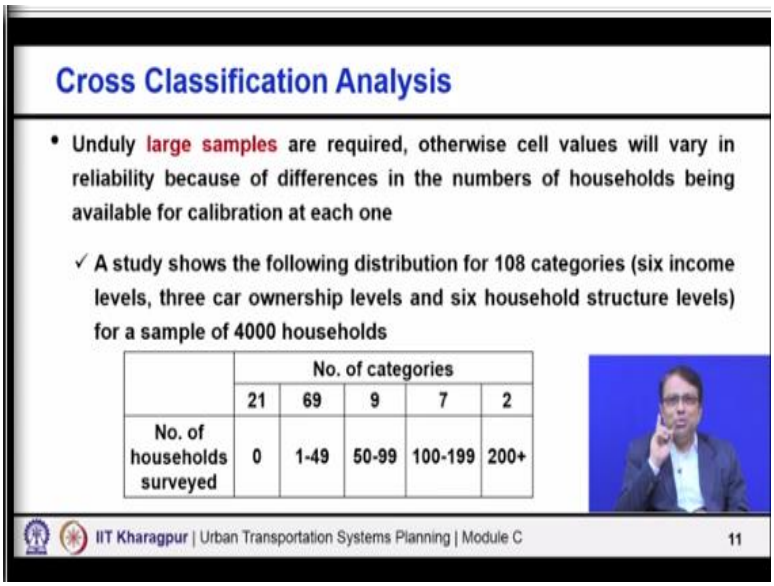
But at least you can see for a given data how good or bad the fit is, you can always reject a model if you are not satisfied with the goodness of fit statistics. And that tells you that gives you confidence, yes, my coefficient estimates are statistically significant, yes, my R square is high which is quite nice. And I know that the model fits to the data well, but here is no statistical goodness of fit.

So, only what you can match is the aggregate prediction, even if you want to apply and try to validate it, if suppose you want to apply it again and validate it, you can only compare the total aggregate thing, that how much I am actually predicting. But no statistical goodness of fit, this is a major weakness because whenever we are working with the data, it means a lot if you can check things statistically.

There are many types of models say for behavioral models, we use stated preference data, reveal preference data, develop behavioral model. In many cases we cannot go and validate those models directly direct validation may not be possible you calculate the WTP value stated preference data, how you will? Will you go to the field and validate it, you cannot do. But then

the statistical goodness of fit, it tells you it gives you good amount of confidence, that is missing here.

(Refer Slide Time: 33:12)



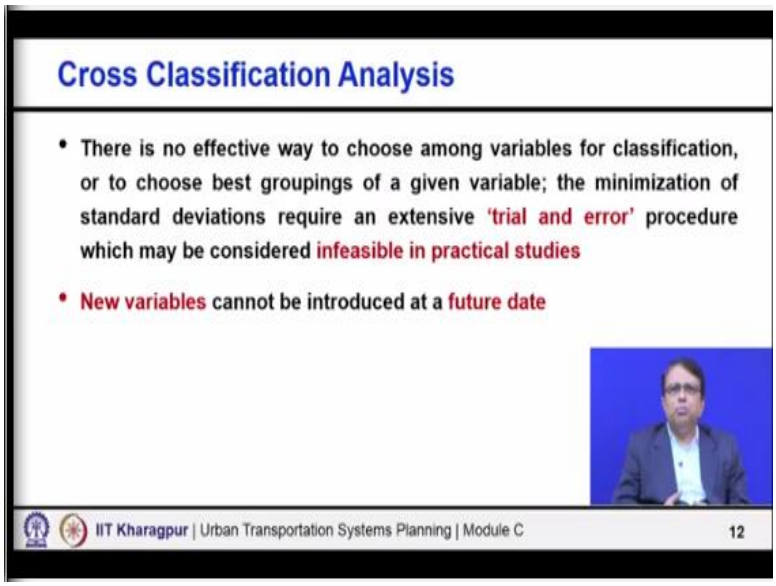
Then unduly large samples are required otherwise cell values will vary in reliability, because of difference in the numbers of household being available in calibration of each one. So, just to tell you this example, as I said one to you every time that you how many cells you are considering, remember that each cell you need representative data. So, this is one interesting example where you say that 4000 households were surveyed, and what they considered?

They consider 6 income levels, 3 car ownership, so 6 into 3, 18 into 6 household structure levels, so into 6 that means 108 categories. So, x, y, z like that income, car ownership and household structure one case is 6 levels, another case 3 level, another case again 6 level. So, 108 categories 4000 household data, what was shown here? There are only 2 cells where 200 plus data were obtained, 7 cases 100 to 199 data were obtained, 9 cases only 50 to 99 data was obtained, 69 cases only 1 to 49 samples were there and 21 cases 0 samples.

See if I simply say that at least 50 samples should be there if we say that. Then only you have representative samples for 9 + 7, 16 + 2, 18 cells out of 108. So, the thing how the it is data hungry, you need to really enormous effort, you need different strategy, different sampling

strategy lot of efforts. So, the requirement of the data is enormous, this is one of the reasons that why many people, planners still use regression based model.

(Refer Slide Time: 35:21)



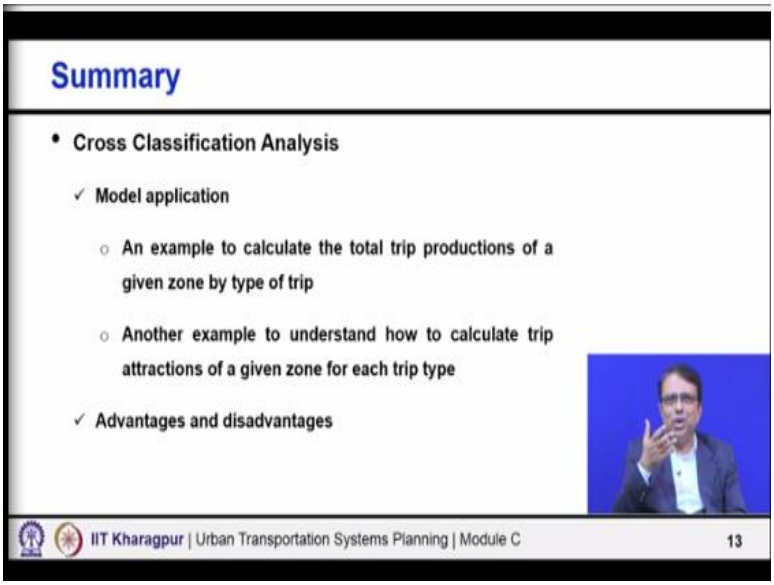
Then there is no effective of choosing among the variables for classification not to choose the best grouping of a given variable. The minimization of a standard deviation require and extensive trial and error procedure which may be considered infeasible in practical studies. Say for example, I said 0 to 20,000, 20,000 to 40,000, 40 to 60 who says that every 20,000 income, this is really that one group is different from other.

Somebody told me that up to 20,000 behavior, travel behavior or trip making behavior does not change, it is only 20,000 beyond it changes, maybe up to 15,000 and 15 to onwards the next group may be different or maybe it is 25,000. We have just rounded it off, take every 20,000 into one group or you may consider, let us say every 25,000 in one groups, these are all arbitrary grouping in a way.

We really do not know, what is the best grouping somewhere the threshold exist, it is not that threshold does not exist, threshold exist. But you do not know what is the base grouping, and by doing trial and error, if you want to find out the grouping that is quite a tedious task. So, most people will not do that, I have classified in 5 groups, I have classified in 3 groups, they will not say on what basis, it is difficult to say.

Last but not the least, new variables cannot be introduced in a future date, if you have considered only income and you know the car ownership, then future also it has to be only income and car ownership, you cannot say that I will consider another variable, you cannot do that. So, those are the advantages and disadvantages. So, you know now the advantages and disadvantages, that helps you to compare these 2 approaches and to select the best one.

(Refer Slide Time: 37:33)



So, what we discussed here? We discussed here 2 things one is basically the model application part, calibration after calibration how to really apply it for the future. Those example 2 examples we have discuss, one for trip production, another for trip attraction. And then we also discussed in details, the advantages and disadvantages associated with cross classification technique or category analysis. So, that you can compare very meaningfully to modeling approaches and decide the best for a given situation. With this I close this lecture, thank you so much.