Trip Generation Model

Introduction

Trip generation is the first stage of the classical first generation aggregate demand models. The trip generation aims at predicting the total number of trips generated and attracted to each zone of the study area. In other words this stage answers the questions to how many trips" originate at each zone, from the data on household and socioeconomic attributes. In this section basic definitions, factors affecting trip generation, and the two main modeling approaches; namely growth factor modeling and regression modeling are discussed.

Types of trip

Some basic definitions are appropriate before we address the classification of trips in detail. We will attempt to clarify the meaning of journey, home based trip, and non-home based trip, trip production, trip attraction and trip generation. Journey is an out way movement from a point of origin to a point of destination, whereas the word trip" denotes an outward and return journey. If either origin or destination of a trip is the home of the trip maker then such trips are called home based trips and the rest of the trips are called non home based trips. Trip production is defined as all the trips of home based or as the origin of the non-home based trips. See figure 1 Trips can be classified by trip purpose, trip time of the day, and by person type. Trip generation models are found to be accurate if separate models are used based on trip purpose. The trips can be classified based on the purpose of the journey as trips for work, trips for education, trips for shopping, trips for recreation and other trips. Among these the work and education trips are often referred as mandatory trips and the rest as discretionary trips. All the above trips are normally home based trips and constitute about 80 to 85 percent of trips. The rest of the trips namely non home based trips, being a small proportion are not normally treated separately. The second way of classification is based on the time of the day when the trips are made. The broad classification is into peak trips and off peak trips. The third way of classification is based on the type of the individual who makes the trips. This is important since the travel behavior is highly influenced by the socio

economic attribute of the traveler and are normally categorized based on the income level, vehicle ownership and house hold size.

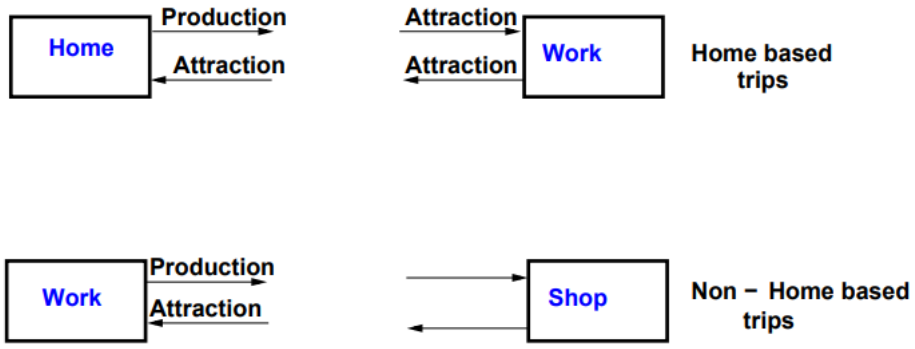

Fig 1: trip types.

Factors affecting trip generation

The main factors affecting personal trip production include income, vehicle ownership, house hold structure and family size. In addition factors like value of land, residential density and accessibility are also considered for modeling at zonal levels. The personal trip attraction, on the other hand, is influenced by factors such as roofed space available for industrial, commercial and other services. At the zonal level zonal employment and accessibility are also used. In trip generation modeling in addition to personal trips, freight trips are also of interest. Although the latter comprises about 20 percent of trips, their contribution to the congestion is significant. Freight trips are influenced by number of employees, number of sales and area of commercial firms.

Trip purpose

It has been found in practice that better trip generation models can be obtained if trips by different purposes are identified and modeled separately. In the case of home-based trips, five categories have been usually employed:

- trips to work;
- trips to school or college (education trips);
- shopping trips;
- social and recreational trips;

- Other trips.

The first two are usually called compulsory (or mandatory) trips and all the others are called discretionary (or optional) trips. The latter category encompasses all trips made for less routine purposes, such as health, bureaucracy (need to obtain a passport or a certificate) and trips made as an accompanying person. Non-home-based trips are normally not separated because they only amount to 15-20% of all trips.

Time of day

Trips are often classified into peak and off-peak period trips; the proportion of journeys by different purposes usually varies greatly with time of day, see Table 1. The morning (AM) peak period (the evening peak period is sometimes assumed to be its mirror image) is usually taken between 7:00 and 9:00 and the representative off-peak period between 10:00 and 12:00.

Table 1: Example of trip classification.

Purpose	AM Peak (%)	Off Peak (%)
work	52.12	12.68
education	35.06	4.96
shopping	1.54	11.35
social	0.79	5.40
health	1.60	2.74
bureaucarcy	3.89	18.35
accompanying	2.09	2.14
other	0.19	0.73
return to home	2.72	41.65

Person type

This is another important classification, as individual travel behavior is heavily dependent on socioeconomic attributes. The following categories are usually employed:

- income level (e.g. nine strata in the Santiago survey);
- car ownership (typically three strata: 0, 1 and 2 or more cars);
- Household size and structure (e.g. six strata in most British studies).

It is important to note that the total number of strata can increase very rapidly and this may have strong implications in terms of data requirements, model calibration and use.

Personal trip productions

The goal of trip production is to estimate the total number of trips, by purpose, produced or originating in each zone. Trip production is performed by relating the number or frequency of trips to the characteristics of the individuals, of the zone, and of the transportation network.

The following factors have been proposed for consideration in many practical studies:

- income;
- car ownership;
- household structure;
- family size;
- value of land;
- residential density;
- Accessibility.

The first four have been considered in several household trip generation studies, while value of land and residential density are typical of zonal studies. The last one, accessibility, has rarely been used although most studies have attempted to include it. The reason is that it offers a way to make trip generation elastic (responsive) to changes in the transport system.

Personal trip attractions

In many ways, estimating trip attractions is similar to estimating trip productions because the problem is the same: predicting the number of trips attracted by relating the number or frequency of trips to the characteristics of the individuals, the zone, and the transportation network. Thus, the methods described in the trip production paragraph -- cross-classification, regression, and discrete choice -- may also be used to estimate the number of trips attracted to a zone. In production models, estimates are primarily based on the demographics of the population within a zone. For attraction models, the variables that have been found to have the best explanatory power are those based on characteristics of the land use, such as office

and retail space or the employment levels of various sectors. As with production models, characteristics of the transportation network are rarely used, which means that the models cannot reflect impacts on trip attractions from changes in accessibility. Also similar to production models, information on the work trip is relatively easy to acquire from such sources as the census or locally initiated surveys. Thus, models of work trip attractions should always be estimated directly using data from the study area, instead of applying models based on national averages or based on another study area. Regression models are often used to estimate trip attractions because of the high correlation between the number of trips made and explanatory variables such as employment and office/retail space (particularly for work trips). Cross-classification can also be used for trip attraction, in which the classification is usually based on the employment sectors, and sometimes employment density. However, the difficulty in collecting the disaggregated data on which to generate the cross-classification table (e.g. it's much easier to collect a statistically valid sample of households than of offices or retail shops) has made crossclassification a rarely used tool for trip attractions. The difficulty of collecting disaggregated data for attraction has also limited the use of discrete choice, although logit models could be applied at the aggregate level.

Freight trip productions and attractions

Freight trips normally account for few vehicular trips; in fact, at most they amount to 20% of all journeys in certain areas of industrialized nations, although they can still be significant in terms of their contribution to congestion. Important variables include:

- number of employees;
- number of sales;
- roofed area of firm;
- Total area of firm.

To our knowledge, neither accessibility nor type of firm have ever been considered as explanatory variables; the latter is curious because it would appear logical that different products should have different transport requirements.

Methods of modeling Trip Generation

1. **Regression Models**: Two types of regression are commonly used. The first uses data aggregated at the zonal level, with average number of trips per household in the zone as the dependent variable and average zonal characteristics as the explanatory variables. The second uses disaggregated data at the household or individual level, with the number of trips made by a household or individual as the dependent variable and the household and personal characteristics as the explanatory variables.
2. **Cross-Classification**: Cross-Classification methods separate the population in an urban area into relatively homogenous groups based on certain socio-economic characteristics. Then, average trip production rates per household or individual are empirically estimated for each class. This creates a lookup table that may be used to forecast trip productions.
3. **Discrete Choice Models**: Discrete choice models use disaggregated household or individual level data to estimate the probability with which any household or individual will make trips. The outcome can then be aggregated to predict the number of trips produced.

Regression methods

Regression methods1 can be used to establish a statistical relationship between the number of trips produced and the characteristics of the individuals, the zone, and the transportation network. Two types of regression models are commonly used. The first uses data aggregated at the zonal level, with average number of trips per household in the zone as the dependent variable and average zonal characteristics as the independent (explanatory) variable. The second uses disaggregated data at the household or individual level, with the number of trips made by a household or individual as the dependent variable and the household and personal characteristics as the independent variables. The best situation is when data for the study area are available that include relevant independent variables (e.g. socio-economic and accessibility factors) and data on frequency of trips for various trip purposes. In this case, you can estimate a regression model that is specifically made for the study area instead of transferring models from another area.

The general form of a trip generation model is

$$T_i = f(x_1, x_2, x_3 \ldots x_i, \ldots x_k) \quad (1)$$

Where x_i's are prediction factor or explanatory variable. The most common form of trip generation model is a linear function of the form

$$T_i = a_0 + a_1x_1 + a_2x_2 + \ldots a_ix_i \ldots + a_kx_k \quad (2)$$

Where: a_i 's are the coefficient of the regression equation and can be obtained by doing regression analysis. The above equations are called multiple linear regression equation, and the solutions are tideous to obtain manually. However for the purpose of illustration, an example with one variable is given.

Example Let the trip rate of a zone is explained by the household size done from the field survey. It was found that the household size are 1, 2, 3 and 4. The trip rates of the corresponding household is as shown in the table below. Fit a linear equation relating trip rate and household size.

Household size(x)				
	1	2	3	4
Trips	1	2	4	6
per	2	4	5	7
day(y)	2	3	3	4
Σy	5	9	12	17

Solution:

The linear equation will have the form y = bx + a where y is the trip rate, and x is the household size, a and b are the coefficients. For a best fit, b is given by:

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$a = \bar{y} - b\bar{x}$$

$$\Sigma x = 3*1 + 3*2 + 3*3 + 3*4 = 30$$

$$\Sigma x^2 = 3*(1^2) + 3*(2^2) + 3*(3^2) + 3*(4^2) = 90$$

$$\Sigma y = 5 + 9 + 12 + 17 = 43$$

$$\Sigma xy = 1*1 + 1*2 + 1*2 + 2*2 + 2*4 + 2*3 + 3*4 + 3*5 + 3*3 + 4*6 + 4*7 + 4*4 = 127$$

$$\bar{y} = 43/12 = 3.58$$

$$\bar{x} = 30/12 = 2.5$$

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = ((12*127) - (30*43))/((12*90) - (30)^2) = 1.3$$

$$a = \bar{y} - b\bar{x} = 3.58 - 1.3*2.5 = -0.33$$

$$\bar{y} = 1.3x + 0.33$$

Zonal-based multiple regression model

In this case an attempt is made to find a linear relationship between the number of trips produced or attracted by zone and average socioeconomic characteristics of the households in each zone. The following are some interesting considerations:

1. Zonal models can only explain the variation in trip making behavior between zones. For this reason they can only be successful if the inter-zonal variations adequately reflect the real reasons behind trip variability. For this to happen it would be necessary that zones not only have a homogeneous socioeconomic composition, but also represent an as wide as possible range of conditions. A major problem is that the main variations in person trip data occur at the intra-zonal level (within zones).
2. Role of the intercept. One would expect the estimated regression line to pass through the origin; however, large intercept values (i.e. in comparison to the product of the average value of any variable and its coefficient) have often been obtained. If this happens the equation may be rejected; if on the contrary, the intercept is not significantly different from zero, it might be informative to re-estimate the line, forcing it to pass through the origin.
3. Null zones. It is possible that certain zones do not offer information about certain dependent variables (e.g. there can be no home based trips generated in non-residential

zones). Null zones must be excluded from analysis; although their inclusion should not greatly affect the coefficient estimates (because the equations should pass through the origin), an arbitrary increment in the number of zones which do not provide useful data will tend to produce statistics which overestimate the accuracy of the estimated regression.

4. Zonal totals versus zonal means. When formulating the model the analyst appears to have a choice between using aggregate or total variables, such as trips per zone and cars per zone, or rates (zonal means), such as trips per household per zone and cars per household per zone. In the first case the regression model would be:

$$T_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + + E_i \qquad 3$$

whereas the model using rates would be:

$$t_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + + e_i \qquad 4$$

with

$t_i = T_i/N_i$; $x_i = X_i/N_i$; $e_i = E_i/N_i$ and N_i the number of households in zone i.

Both equations are identical, in the sense that they seek to explain the variability of trip making behavior between zones, and in both cases the parameters have the same meaning. Their unique and fundamental difference relates to the error-term distribution in each case; it is obvious that the constant variance condition of the model cannot hold in both cases, unless Ni was itself constant for all zones i.

Now, as the aggregate variables directly reflect the size of the zone, their use should imply that the magnitude of the error actually depends on zone size; this heteroscedasticity (variability of the variance) has indeed been found in practice. Using multipliers, such as 1/Ni, allows heteroscedasticity to be reduced because the model is made independent of zone size. In this same vein, it has also been found that the aggregate variables tend to have higher intercorrelation (i.e. multicollinearity) than the mean variables. However, it is important to note that models using aggregate variables often yield higher values of R^2, but this is just a spurious effect because zone

size obviously helps to explain the total number of trips. What is certainly unsound is the mixture of means and aggregate variables in a single model.

The various difficulties encountered with zonal regression models (dependence on zone size, zonal boundaries, spurious correlations, etc.) have led to the use of models based on the true behavioral units: households or persons.

Household-based regression model

Intra-zonal variation may be reduced by decreasing zone size, especially if zones are homogeneous. However, smaller zones imply a greater number of them and this has two consequences:

- more expensive models in terms of data collection, calibration and operation;
- greater sampling errors, which are assumed non-existent by the multiple linear regression model.

For these reasons it seems logical to postulate models which are independent of zonal boundaries. At the beginning of the 1970s it was decided that the most appropriate analysis unit in this case was the household (and not the individual); it was argued that a series of important interpersonal interactions inside a household could not be incorporated even implicitly in an individual model (e.g. car availability, that is, who has use of the car).

In a household-based application each home is taken as an input data vector in order to bring into the model all the range of observed variability about the characteristics of the household and its travel behavior. The calibration process, as in the case of zonal models, proceeds stepwise, testing each variable in turn until the best model (in terms of some summary statistics for a given confidence level) is obtained. Care has to be taken with automatic stepwise computer packages because they may leave out variables which are slightly worse predictors than others left in the model, but which may prove much easier to forecast.

Example 1: Consider the variables trips per household (t), number of workers (X_1) and number of cars (X_2). Table 2 presents the results of successive steps of a step-wise model estimation. Assuming a large sample size, the appropriate number of degrees of freedom (n – 2) is also a large number so the t-values may be compared with the critical value 1.645 for a 95% significance level on a one-tailed test (we know the null hypothesis is unilateral in this case as t should increase with both X_1 and X_2). The third model is a good equation in spite of its low R^2. The intercept 0.91 is not large (compare it with 1.44 times the number of workers, for example) and the explanatory variables are significantly different from zero (N_0 is rejected in all cases). The model could probably benefit from the inclusion of other variables.

Table 2: Example of Stepwise Regression

Step	Equation	R^2
1	$t = 2.36\ X_1$	0.203
2	$t = 1.80\ X_1 + 1.31\ X_2$	0.325
3	$t = 0.91 + 1.44\ X_1 + 1.07\ X_2$	0.384

No. of cars	Number of workers in household			
	0	1	2	3 or more
0	0.9/0.9	2.1/2.4	3.4/3.8	5.3/5.6
1	3.2/2.0	3.5/3.4	3.7/4.9	8.5/6.7
2 or more	-	4.1/4.6	4.7/6.0	8.5/7.8

An indication of how good these models are may be obtained from comparing observed and modeled trips for some groupings of the data (see Table 2). This is better than comparing totals because in such case different errors may compensate and the bias would not be detected. As can be seen, the majority of cells show a reasonable approximation (i.e. errors of less than 30%). If large biases were spotted it would be necessary to adjust the model parameters; however, this is not easy as there are no clear-cut rules to do it, and it depends heavily on context.

Example 2: The following trip frequency model (see Table 3) is an example of a household-based trip production model. It is a linear regression model using dummy variables. It calculates the weekly number of trips by individual households. The level of (weekly) trip making appears to depend on:

- size of the household
- life cycle
- highest education in the household
- structure of the household
- Number of driving license owners Income and car ownership appear not to be significant as explanatory variables.

This is most probably a result of the unspecific nature of the models that is no distinction between trip purposes. The model has been estimated using data from the Dutch Longitudinal Mobility Panel [see Bovy & Kitamura, 1986]. On average, a Dutch household makes about 50 trips a week. The largest contribution is given by the number of household members: 22.5 trips per person per week. The higher the educational level the more trips are made with in the extreme a difference of 14.1 trips per week (10.1– – 4.0). The model performs extremely well with it explained trip variance of 90%.

Model:

$$Y = \sum_k \propto_k X_k \qquad 5$$

Y: weekly # household trips (all purposes, all modes).

Table 3 Regression model (with dummy variable) of household weekly trip production (Source: Bovy & Kitamura, 1986).

Variable		Parameter value
# members household		22.5
lifestyle	couple, no kids, male < 35 year	5.6
	couple, no kids, male 35-64 year	-2.8
	couple, no kids, male > 64 year	-5.6
education	elementary school	-4
	intermediate vocational or higher general preparatory school	3.4
	higher vocational school	9.4
	university	10.1
# driver license		3.9
structure	# adults	-5
	# children < 6 year	1.9
	# children 6-12 year	3.5
Goodness-of-fit $R^2 = 0.9$		

The problem of non-linearities

As we have seen, the linear regression model assumes that each independent variable exerts a linear influence on the dependent variable. It is not easy to detect non-linearity because apparently linear relations may turn out to be non-linear when the presence of other variables is allowed in the model. Multivariate graphs are useful in this sense; the example of Figure 2 presents data for households stratified by car ownership and number of workers. It can be seen that travel behavior is non-linear with respect to family size.

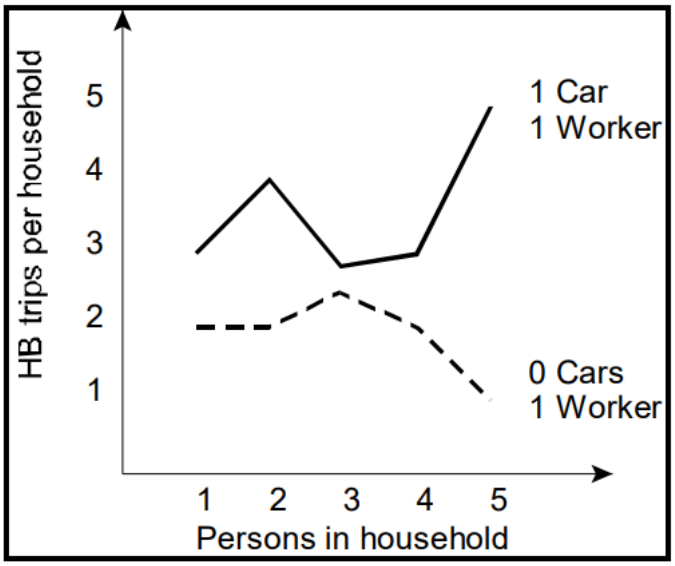

Figure 2: An example of non-linearity.

It is important to mention that there is a class of variables, those of a qualitative nature, which usually shows non-linear behavior (e.g. type of dwelling, occupation of the head of the household, age, sex). In general there are two methods to incorporate non-linear variables into the model:

1. Transform the variables in order to linearize their effect (e.g. take logarithms, raise to a power). However, selecting the most adequate transformation is not an easy or arbitrary exercise, so care is needed; also, if we are thorough, it can take a lot of time and effort;
2. Use dummy variables. In this case the independent variable under consideration is divided into several discrete intervals and each of them is treated separately in the model.

In this form it is not necessary to assume that the variable has a linear effect, because each of its portions is considered separately in terms of its effect on travel behavior. For example, if car ownership was treated in this way, appropriate intervals could be 0, 1 and 2 or more cars per household. As each sampled household can only belong to one of the intervals, the corresponding dummy variable takes a value of 1 in that class and 0 in the others. It is easy to see that only (n - 1) dummy variables are needed to represent n intervals.

Example 3: Consider the model of Example 1 and assume that variable X_2 is replaced by the following dummies: Z_1, which takes the value 1 for households with one car and 0 in other cases; Z_2, which takes the value 1 for households with two or more cars and 0 in other cases. It is easy to see that non-car-owning households correspond to the case where both Z_1 and Z_2 are 0. The model of the third step in Table 2 would now be:

$$\boldsymbol{t = 0.84 + 1.41X_1 + 0.75Z_1 + 3.14\,Z_2} \qquad \textbf{4}$$

R^2=0.387

Even without the better R^2 value, this model is preferable to the previous one just because the nonlinear effect of X_2 (or Z_1 and Z_2) is clearly evident and cannot be ignored. Note that if the coefficients of the dummy variables were for example, 1 and 2, and if the sample never contained more than two cars per household, the effect would be clearly linear. The model is graphically depicted in Figure 3.

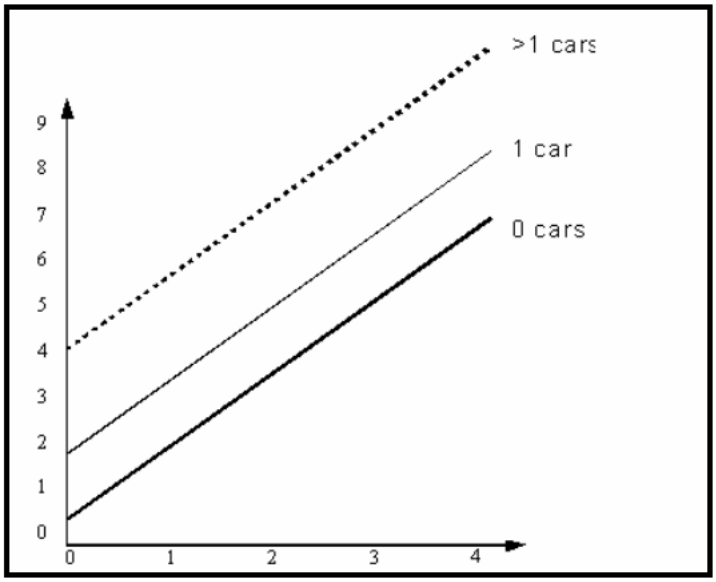

Figure 3: Regression model with dummy variables.

Cross-classification or category analysis model

Cross-classification methods of calculating productions separate the population in an urban area into relatively homogenous groups (households or persons) based on certain socioeconomic characteristics. For example, one may classify households in an area by both family size (1, 2, 3, 4, ≥ 5 persons/household) and by auto ownership (0, 1, ≥ 2 autos/household), which results in 15 classes (see Table 4). Average trip-production rates (the estimated number of trips that will be taken by a household or individual) are empirically derived from either disaggregated or aggregate data sets for each of the classes. In the example above, 15 average trip rates would be derived.

Once trip rates are known for each class, these trip rates are usually applied to each zone.

Table 4: Example Cross-Classification Table.

Family Size	0 cars	1 car	≥2 cars
1			
2			
3			
4			
≥ 5			

Each zone may be subdivided into a few classes by using the proportion of households or persons within a zone that have a certain characteristic. Using this method, more than one average trip rate is used to estimate productions for any one zone. For example, a zone may be divided into households without cars and households with cars. In this case, 2 average trip rates will be applied to each zone.

$$T_i^p = \sum_h N_{hi}\, t_h^p \qquad 5$$

The household-based category model

The method is based on estimating the response (e.g. the number of trip productions per household for a given purpose) as a function of household attributes. Its basic assumption is that trip generation rates are relatively stable over time for certain household stratifications. The method finds these rates empirically and for this it typically needs a large amount of data; in fact, a critical element is the number of households in each class.

Let t_h^p be the average number of trips with purpose p (and at a certain time period) made by members of households of type h. Types are defined by the stratification chosen; for example, a cross-classification based on m household sizes and n car ownership classes will yield mn types h. The standard method for computing these cell rates is to allocate households in the calibration data to the individual cell groupings and total, cell by cell, the observed trips T_h^p by purpose group. The rate t_h^p is then the total number of trips in cell h, by purpose, divided by the number of households N_h in it. In mathematical form it is simply:

$$t_h^p = \frac{T_h^{\acute{p}}}{N_h} \qquad 6$$

The 'art' of the method lies in choosing the categories such that the standard deviations of the frequency distributions depicted in Figure 4 are minimized.

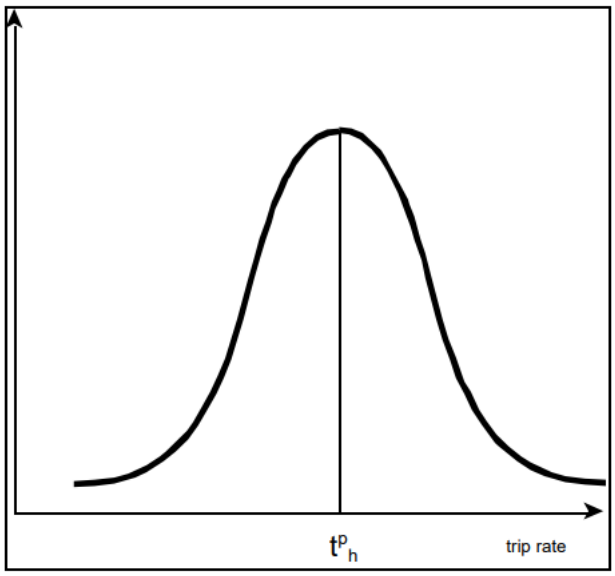

Figure 4: Trip-rate distribution for household type.

The method has, in principle, the following advantages:

1. Cross-classification groupings are independent of the zone system of the study area.
2. No prior assumptions about the shape of the relationship are required (i.e. they do not even have to be monotonous, let alone linear).
3. Relationships can differ in form from class to class (e.g. the effect of changes in household size for one or two car-owning households may be different).

And in common with traditional cross-classification methods it has also several disadvantages:

1. The model does not permit extrapolation beyond its calibration strata, although the lowest or highest class of a variable may be open-ended (e.g. households with two or more cars and five or more residents).
2. Unduly large samples are required, otherwise cell values will vary in reliability because of differences in the numbers of households being available for calibration at each one. Accepted wisdom suggests that at least 50 observations per cell are required to estimate the mean reliably.
3. There is no effective way to choose among variables for classification, or to choose best groupings of a given variable; the minimization of standard deviations hinted at in Figure 4 requires an extensive 'trial and error' procedure which may be considered infeasible in practical studies.

Model application at zonal level

Let us denote by h the household type (i.e. with and without a car), be N_{hi} the number of households of type h in zone i. With this we can write the trip productions with purpose p by household type h in zone i, T_h^p, as follows in Eq. 5:

$$T_i^p = \sum_h N_{hi}\, t_h^p$$

To verify how the model works it is possible to compare these modeled values with observed values from the calibration sample. Inevitable errors are due to the use of averages for thet_h^p; one would expect a better stratification (in the sense of minimizing the standard deviation in Figure 4) to produce smaller errors.

There are various ways of defining household categories. The first application in the UK (Wootton and Pick 1967), which was followed closely by subsequent applications, employed 108 categories as follows: six income levels, three car ownership levels (0, 1 and 2 or more cars per household) and six household structure groupings, as in Table 5. The problem is clearly how to predict the number of households in each category in the future.

Table 5: Example of household structure grouping.

Group	No. employed	Other adults
1	0	1
2	0	2 or more
3	1	1 or less
4	1	2 or more
5	2 or more	1 or less
6	2 or more	2 or more

The following Table 6 is an example of a personal trip generation model consisting of a production and an attraction part. The table includes trip rates per unit for the evening peak hour. Separate models are given for three spatial settings (agglomerations, towns and rural) and for two trip purposes each, namely home-to-work and other.

Table 5: Personal trip rates used in the Randstadmodel [Source: Randstadmodel 1994].

evening peak			production		attraction		total
			labour-force	inhabitant	employment		
					retail	other	
agglomeration	home-work	arrivals	0.2282				∑
		departures				0.2626	∑
	other	arrivals		0.1259	0.3224	0.0257	∑
		departures		0.0633	0.3173	0.2626	∑
town	home-work	arrivals	0.2169				∑
		departures				0.2435	∑
	other	arrivals		0.1633	0.5934	0.0323	∑
		departures		0.895	0.6127	0.2435	∑
rural	home-work	arrivals	0.236				∑
		departures				0.2674	∑
	other	arrivals		0.13	0.8417	0.0372	∑
		departures		0.0744	0.7039	0.2674	∑

Production trip rates for home-to-work are based on working persons, whereas all other production trip rates refer to inhabitants. Trip attractions are based on employment by type (retail and other).

The trip rates were estimated using various data sources (OVG). Calculation of total origins and destinations per zone is performed by horizontally adding the various production and attraction components.

Estimation of trip rates by multiple class analysis (MCA)

MCA is an alternative method to define classes and test the resulting cross-classification which provides a statistically powerful procedure for variable selection and classification. This allows us to overcome several of the disadvantages cited above for other types of cross-classification methods

Consider a model with a continuous dependent variable (such as the trip rate) and two discrete independent variables, such as household size and car ownership. A grand mean can be estimated for the dependent variable over the entire sample of households. Also, group means can be estimated for each row and column of the cross-classification matrix; each of these can be expressed in turn as deviations from the grand mean. Observing the signs of the

deviations, a cell value can be estimated by adding to the grand mean the row and column deviations corresponding to the cell. In this way, some of the problems arising from too few observations on some cells can be compensated.

Example 4: Table 6 presents data collected in a study area and classified by three car-ownership and four household-size levels. The table presents the number of households observed in each cell (category) and the mean number of trips calculated over rows, cells and the grand average.

Table 5: Number of households per cell and mean trip rates for a particular purpose.

Household size	0 car	1 car	2+ cars	Total	Mean
1 person	28	21	0	49	0.47
2 or 3 persons	150	201	93	444	1.28
4 persons	61	90	75	226	1.86
5 persons	37	142	90	269	1.9
Total	276	454	258	988	
Mean trip rate	0.73	1.53	2.44		1.54 (= grand mean)

As can be seen, the values range from 0 (it is unlikely to find households with one person and more than one car) to 201. Although we are cross-classifying by only two variables in this simple example, there are already four cells with less than the conventional minimum number (50) of observations required to estimate mean trip rate and variance with some reliability. We would like to use now the mean row and column values to estimate average trip rates for each cell, including that without observations in this sample.

We can compute the deviation (from the grand mean) for zero cars as 0.73 - 1.54 = -0.81; for one car as 1.53 - 1.54 = -0.01, and for two cars or more 2.44 - 1.54 = 0.90; similarly, we can calculate the deviations for each of the four household size groups as: -1.07, -0.26, 0.32 and 0.36. If the variables are not correlated with these values we can work out the full trip-rate Table 6; for example, the trip rate for one person household and one car is 1.54 - 1.07 - 0.01 = 0.46 trips. In the case of one person and no car, the rate turns out to be negative and equal to -0.34 (1.54 -1.06- 0.82); this has no meaning and therefore the actual rate is forced to zero.

Table 6: depicts the full trip-rate table together with its row and column deviations.

Household size	Car ownership level			
	0 car	1 car	2+ cars	Deviations from grand mean
1 person	0	0.46	1.37	-1.07
2 or 3 persons	0.46	1.27	2.18	-0.26
4 persons	1.05	1.85	2.76	0.32
5 persons	1.09	1.89	2.8	0.36
Deviations	-0.81	-0.01	0.9	

Contrary to standard cross-classification models, deviations are not only computed for households in, say, the cell one person-one car; rather, car deviations are computed over all household sizes and vice versa. Thus, if interactions are present these deviations should be adjusted to account for interaction effects. This can be done by taking a weighted mean for each of the group means of one independent variable over the groupings of the other independent variables, rather than a simple mean (which would in fact be equivalent to assuming that variation is random over the data in a group). These weighted means will in general tend to decrease the sizes of the adjustments to the grand mean when interactions are present. Nevertheless, the cell means of a multiway classification will still be based on means estimated from all the available data, rather than being based on only those items of data falling in the multiway cell.

Apart from the statistical advantages, it is important to note that cell values are no longer based on only the size of the data sample within a given cell; rather, they are based on a grand mean derived from the entire data set, and on two (or more) class means which are derived from all data in each class relevant to the cell in question.

Example 5: Table 6 provides a set of rates computed in the standard category analysis procedure (i.e. by using individual cell means). These values may be compared with those of Table 7. Two points of interest emerge from the comparison. First, there are rates available even for empty cells in the MCA case. Second, some counterintuitive progressions, apparent in Table 7 (e.g. the decrease of rate values for 0 and 1 car-owning households when increasing household size from 4 to 5 or more), are removed in Table 7. Note that they could have arisen by problems of small sample size at least in one case.

Table 7: Trip rates calculated using ordinary category analysis.

Household size	Car ownership level		
	0 car	1 car	2+ cars
1 person	0.12	0.94	
2 or 3 persons	0.6	1.38	2.16
4 persons	1.14	1.74	2.6
5 persons	1.02	1.69	2.6

The person-category approach

This is an interesting alternative to the household-based models discussed above. This approach offers the following advantages:

1. A person-level trip generation model is compatible with other components of the classical transport demand modeling system, which is based on trip makers rather than on households; 2. It allows a cross-classification scheme that uses all important variables and yields a manageable number of classes; this in turn allows class representation to be forecast more easily;
2. The sample size required to develop a person-category model can be several times smaller than that required to estimate a household-category model;
3. Demographic changes can be more easily accounted for in a person-category model as, for example, certain key demographic variables (such as age) are virtually impossible to define at household level;
4. Person categories are easier to forecast than household categories as the latter require forecasts about household formation and family size; these tasks are altogether avoided in the case of person categories. In general the bulk of the trips are made by people older than 18 years of age; this population is easier to forecast 15 to 20 years ahead as only migration and survival rates are needed to do so.

The major limitation that a person-category model may have relates precisely to the main reason why household-based models were chosen to replace zonal-based models at the end of the 1960s; this is the difficulty of introducing household interaction effects and household money costs and money budgets into a person-based model.

Variable definition and model specification

The estimation of person-based trip rates per person type follows the same line as explained before with respect to households (MCA). Model development entails the following stages:

1. Consideration of several variables which are expected to be important for explaining differences in personal mobility. Also, definition of plausible person categories using these variables;
2. Preliminary analysis of trip rates in order to find out which variables have the least explanatory power and can be excluded from the model. This is done by comparing the trip rates of categories which are differentiated by the analyzed variable only and testing whether their differences are statistically significant;
3. Detailed analysis of trip characteristics to find variables that define similar categories. Variables which do not provide substantial explanation of the data variance, or variables that duplicate the explanation provided by other better variables (i.e. easier to forecast or more policy responsive) are excluded. The exercise is conducted under the constraint that the number of final categories should not exceed a certain practical maximum (for example, 15 classes).

For this analysis the following measures may be used: the coefficient of correlation R_{jk}, slope m_{jk} and intercept a_{jk} of the regression.

$$\mathbf{t_h^p = a_{jk} + m_{jk} t_k^p}$$

The categories j and k may be treated as similar if these measures satisfy the following conditions:

$R_{jk} > 0.900$

$0.75 < m_{jk} < 1.25$

$a_{jk} < 0.10$

These conditions are quite demanding and may be changed.

Model application at the aggregate level

Let t_n be the trip rate, that is, the number of trips made during a certain time period by (the average) person in category j; t_h^p, is the trip rate by purpose p. T_i is the total number of trips made by the inhabitants of zone i (all categories together). N_i is the number of inhabitants of zone i, and α_{ni} is the percentage of inhabitants of zone i belonging to category n. Therefore the following basic relationship exists:

$$T_i^p = N_i \sum_n \alpha_{ni} t_h^p \qquad 6$$

Example 6: Another Dutch example of a cross-classification trip rate model is WOLOCAS. This model predicts trip productions for new residential areas. The trip rates refer to person types classified by sex, car ownership, occupational status, education and age (see Tables 8-10). The trip production model distinguishes three trip purposes. The daily trip rates were estimated using the continuous National Dutch Mobility Survey. In applying this model estimates are required for the demographic size and composition of the new residential areas. The trip rates only apply to persons over 12 years old. [Source: Wolocas, 1990].

Table 8: Personal daily trip production rates for "work", split by person type.

Trip purpose WORK		education		
		low	medium	HBO/univ
working man	with car	1.7	1.743	1.702
	without car	1.656	1.656	1.656
working woman	with car	1.369	1.331	1.331
	without car	1.268	1.31	1.31

Table 9: Personal daily trip production rates for "services", split by person type.

Trip purpose SERVICES		age	
		12 - 18 year	> 18 year
man	with car		0.656
	without car	1.983	0.917
woman	with car	-	1.111
	without car	1.983	1.201
average		1.983	

Table 10: Personal daily trip production rates for "other", split by person type.

Trip purpose OTHER		age	
		12 – 40 year	> 40 year
working	with car	1.185	0.779
	without car	0.974	0.98
not-working	with car	1.158	1.442
	without car	1.512	0.98

Discrete choice methods

Since individuals choose whether to make specific trips, discrete choice models such as binary logit can be used to predict trip production. With binary logit, the probability that an individual will choose to make one or more trips (as opposed to not traveling) can be expressed as:

$$P_n(+1) = \frac{1}{1 + e^{\beta(x_{0n} + x_{1n}}}$$

$$P_n(0) = 1 - P_n(+1) \qquad 7$$

Where:

Pn(0) = the probability that a person n will make no trip
Pn(1+) = the probability that a person n will make one or more trips
β = the vector of coefficients that is estimated by the model
x_{1n} = the vector of explanatory variables in person n's utility of making one or more trips
x_{0n} = the vector of explanatory variables in person n's utility of not making a trip

From the estimated coefficients, you can see how the explanatory variables will impact the probability with which an individual will make a certain trip. In addition, you can aggregate the disaggregated probabilities to obtain the proportion of the population that will take this type of trip, and thus generate the aggregate number of trips produced by a zone.

Interpreting the results of a logit model

A model yielded the following results:

Table 11: Results of a logit model predicting trip making probability

Parameter	Estimate	t-Stat
Constant	-0.474	-2.4
Sex	0.267	2.1
Age	-0.047	-19.1
Married Female	0.314	2.8
Married Male	1.594	12.4
Fem w/ Child<6	-1.742	-11.4
Education	0.211	13.8

With a goodness-of-fit ('Adjusted Rho Squared') of 0.22. The variables Sex, Age, Married Female, Married Male, Fem w/ Child < 6 are all dummy variables (i.e. equal to 1 or 0). Note that all of the coefficients are significant at a 95% confidence level (t-statistic > 2).

This logit model predicts the probability with which an individual will make a work trip according to the following equation:

$$P_{Trip} = \frac{1}{1 + e^{0.474 - 0.267(sex) + 0.047(Age) - 0.314(MarFem) - 1.594(MarMale) + 1.742(Chld<6) - 0.211(Educ)}}$$

$$P_{no\ trip} = 1 - P(trip) \qquad (8)$$

From the estimated coefficients, you can see how the explanatory variables will impact the probability with which an individual makes a trip to work. For example, the coefficient for education (0.211) suggests that, all else being equal, people with more education are more likely to make work trips than those with less education. Note that none of the signs from the model seem unreasonable. You can also use the equation above to calculate how a change in an explanatory variable will impact a person's probability of making a work trip. For example, a person with a high school diploma (Educ = 10) who has a 50% probability of making a work trip, would, all else being equal, have a 70% probability of making a work trip if he or she had a Bachelor's Degree.

Example 7: The Dutch National Transport Model System is fully based on disaggregated discrete choice models of travel demand. The trip frequency submodel predicts the probability that an individual with known characteristics will make no, one or more round tours for specific trip purposes on an average day. In later stages, tours are decomposed into trips. The unit of analysis is a person, member of a household with known characteristics. This trip production model distinguishes 5 trip purposes, further divided by person type. In total, eight different models are distinguished.

The model has a two-stage structure:

- in step 1, the model predicts the probability of making no versus one or more tours, thus a binary choice model.
- in step 2, another binary choice model predicts the conditional probability of making one tour versus two or more tours given the fact that tours are being made. This process is repeated several times.

Model 2 is applied consecutively in the cascade. This second model is identical in all these steps (but is different from model 1). It is a so-called stop/repeat-model. Both models are binary logit models with linear utility functions including dummy variables describing personal and household characteristics. In practical applications, instead of working with probabilities, the expected number of tours is calculated.

This expected number of tours (ENT) per purpose may be calculated using:

$$\boldsymbol{ENT = \frac{P_r(+1)}{1-P_r(R)}} \qquad \mathbf{9}$$

Where $P_r(1+)$ is the probability outcome of model 1, and $P_r(R)$ is the probability of making additional trips which is the outcome of model 2.

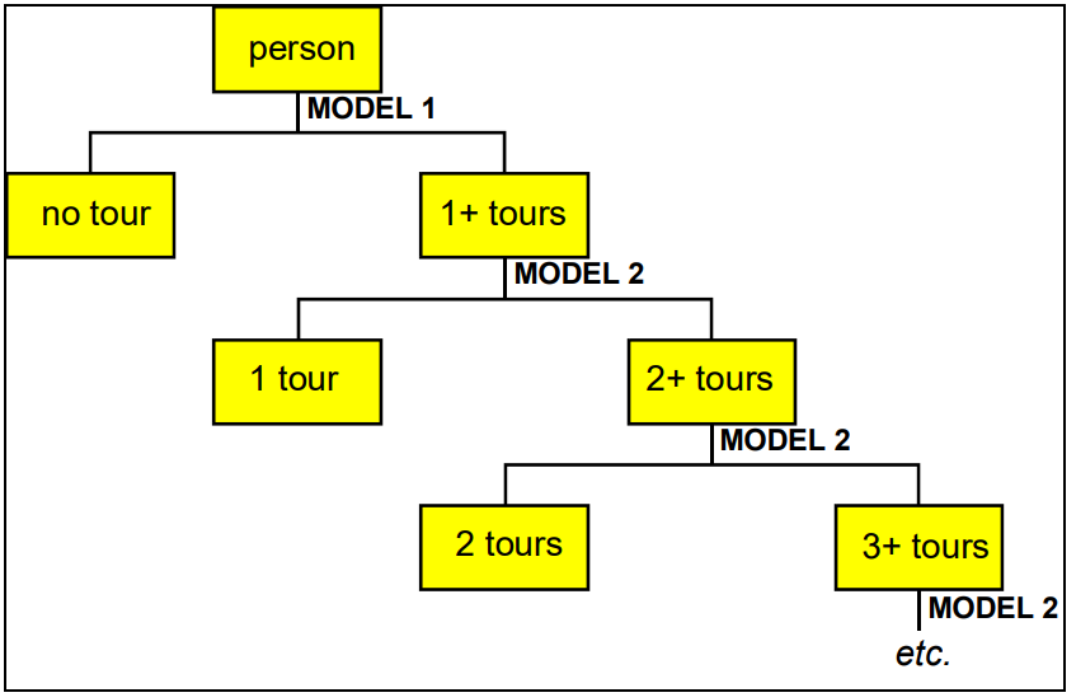

Figure 5: Structure of disaggregated trip choice models for trip production [Source: LMS, 1996]

Trip balancing

It might be obvious to some readers that the models above do not guarantee, by default, that the total number of trips originating (the origins O_i) at all zones will be equal to the total number of trips attracted (the destinations D_j) to them, that is the following expression does not necessarily hold:

$$\sum_i O_i = \sum_j D_j \qquad 10$$

The problem is that this equation is implicitly required by the next sub-model (i.e. trip distribution) in the structure; it is not possible to have a trip distribution matrix where the total number of trips (T) obtained by summing all rows is different to that obtained when summing all columns.

The solution to this difficulty is a pragmatic one which takes advantage of the fact that normally the trip generation models are far 'better' (in every sense of the word) than their trip attraction counterparts. The first normally are fairly sophisticated household-based models with typically good explanatory variables. The trip attraction models, on the other hand, are

at best estimated using zonal data. For this reason, normal practice considers that the total number of trips arising from summing all origins O_i is in fact the correct figure for T; therefore, all destinations D_j are multiplied by a factor f given by:

$$f = \frac{T}{\sum_j D_j} \qquad 11$$

Which obviously ensure that their sum adds to T.

Several procedures can be used to balance trip productions and attractions in which productions and attractions from several trip purposes can be balanced in one step. The procedure offers the following methods for balancing:

- Hold Productions Constant Productions are held constant and the attractions are adjusted so that their sum equals the sum of the productions.
- Hold Attractions Constant Attractions are held constant and the productions are adjusted so that their sum equals the sum of the attractions.
- Weighted Sum of Productions and Attractions Both productions and attractions are adjusted so that their sums equal the user specified weighted sum of productions and attractions.
- Sum to User Specified Value Both productions and attractions are adjusted so that their sums equal a user specified value.

Table 12 illustrated model specifications for trip generation models.

Table 12: Summary of model specifications for trip generation models.

REGRESSION MODELS		
level	**equation**	**notes**
zonal	$T_i^p = \sum_k \alpha_k^p X_{ik}$	T_{ip} = zonal production of trips for purpose p X_{ik} = k^{th} zonal explanatory variable of zone i
household	$T_i^p = \sum_h N_{hi} \sum_k \beta_{hk}^p Y_{hk}$	Y_{hk} = k^{th} household explanatory variable for household type h N_{hi} = no. of households of type h in zone i
person	$T_i^p = \sum_n N_{ni} \sum_k \gamma_{nk}^p Z_{nk}$	Z_{pk} = k^{th} personal explanatory variable for person type p N_{ni} = no. of persons of type n in zone i
CROSS CLASSIFICATION		
level	**equation**	**notes**
household	$T_i^p = \sum_h N_{hi} t_h^p$	t_h^p = trip rate per household of type h
person	$T_i^p = \sum_n N_{ni} t_n^p$	t_n^p = trip rate per person of type n
DISCRETE CHOICE MODELS		
level	**equation**	**notes**
household	$T_i^p = \sum_h N_{hi} t_h^p \sum_{x=0}^{\max} x P_{hp}(x)$	$P_{hp}(x)$ = probability that a random household from householdgroup h will make x trips (x = 0,1,2,..,max) for purpose p
person	$T_i^p = \sum_n N_{ni} t_n^p \sum_{x=0}^{\max} x P_{np}(x)$	$P_{np}(x)$ = probability that a random person from persongroup n will make x trips (x = 0,1,..,max) for purpose p

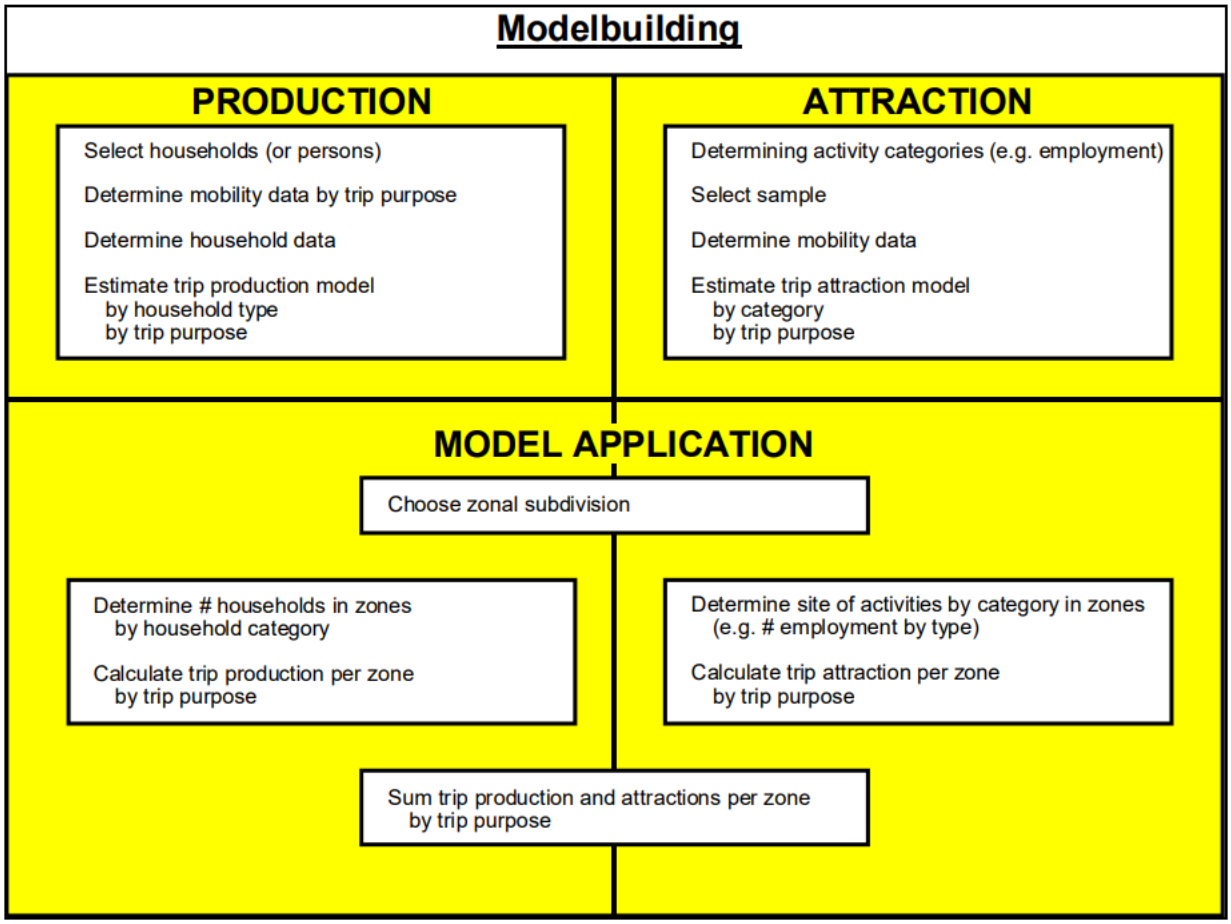

Figure 13: Summary of procedure to calculate origins (departure) and zonal destinations (arrivals).