Choice-based sample

<table>
<thead>
<tr>
<th>Mode</th>
<th>Low I</th>
<th>Medium I</th>
<th>High I</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>6.67</td>
<td>20.00</td>
<td>13.33</td>
<td>40.00</td>
</tr>
<tr>
<td>Bus</td>
<td>17.24</td>
<td>9.07</td>
<td>0.69</td>
<td>20.00</td>
</tr>
<tr>
<td>Underground</td>
<td>16.67</td>
<td>13.33</td>
<td>10.00</td>
<td>40.00</td>
</tr>
</tbody>
</table>

(a) If you know that the income-based proportions in the population are 60, 25 and 15% respectively for low, medium and high income, find an equivalent table for a random sample. Is it possible to validate your answer?

(b) Compute the weighting factors that would be necessary to apply to the observations in the choice-based sample in order to estimate a model for the choice between car, bus and underground using standard software (i.e. that developed for random samples).

4 Trip Generation Modelling

This is the first of several chapters dealing with first-generation aggregate demand models. As we saw in Chapter 1, the trip generation stage of the classical transport model aims at predicting the total number of trips generated by ($O_i$) and attracted to ($D_j$) each zone of the study area. This has been usually considered as the problem of answering a question such as: how many trips originate at each zone? However, the subject has also been viewed sometimes as a trip frequency choice problem: how many shopping (or other purpose) trips will be carried out by this person type during an average week? This is usually undertaken using discrete choice models, as discussed in Chapters 7 to 9, and it is then cast in terms like: what is the probability that this person type will undertake zero, one, two or more trips with this purpose per week?

In this chapter we concentrate on the first approach (i.e. predicting the totals $O_i$ and $D_j$ from data on household socioeconomic attributes), which has been the most widely used in practice up to the end of the 1990s. Readers interested in the discrete choice approach should consult Ben Akiva and Lerman (1985) and, of course, Chapters 7 to 9 in this book.

We will start by defining some basic concepts and will proceed to examine some of the factors affecting the generation and attraction of trips. Then we will review the main modelling approaches, starting with the simplest growth-factor technique. Before embarking on more sophisticated approaches we will present a reasonable review of linear regression modelling, which complements well the previous statistical themes presented in Chapters 2 and 3.

We will then consider zonal and household-based linear regression trip generation models, giving some emphasis to the problem of non-linearities which often arise in this case. We will also address for the first time the problem of aggregation (e.g. obtaining zonal totals), which has a trivial solution here precisely because of the linear form of the model. Then we will move to cross-classification models, where we will examine not only the classical category analysis specification but also more contemporary approaches including the person category analysis model. We then examine the relationship between trip generation and accessibility including a short discussion on trip frequency models. The chapter ends with two short sections: the first discusses the problem of predicting future values for the explanatory variables in the models, and the second the problems of stability and updating of trip generation parameters.
4.1 INTRODUCTION

4.1.1 Some Basic Definitions

**Journey** This is a one-way movement from a point of origin to a point of destination. Now, although the word ‘trip’ is literally defined as ‘an outward and return journey’, often for a specific purpose (McLeod and Hanks 1986), in transport modelling both terms are used interchangeably. We are usually interested in all vehicular trips, but walking trips longer than a certain study-defined threshold (say 300 metres or three blocks) are often considered; finally, trips made by infants of less than five years of age will usually be ignored.

**Home-based (HB) Trip** This is one where the home of the trip maker is either the origin or the destination of the journey.

**Non-home-based (NHB) Trip** This, conversely, is one where neither end of the trip is the home of the traveller.

**Trip Production** This is defined as the home end of an HB trip or as the origin of an NHB trip (see Figure 4.1).

**Trip Attraction** This is defined as the non-home end of an HB trip or the destination of an NHB trip (see Figure 4.1).

**Trip Generation** This is often defined as the total number of trips generated by households in a zone, be they HB or NHB. This is what most models would produce and the task then remains of allocating NHB trips to other zones as trip productions.

During the 1980s a series of other terms, such as sojourns, tours and trip chains, were added to the transport modelling kit; these correspond better to the idea that the demand for travel is a derived demand (i.e. it depends strongly on the demand for other activities) but have been used mainly by discrete choice modellers in practice (see Daly et al. 1983).

4.1.2 Classification of Trips

4.1.2.1 By Trip Purpose

It has been found in practice that better trip generation models can be obtained if trips by different purposes are identified and modelled separately. In the case of HB 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. NHB trips are normally not separated because they only amount to 15–20% of all trips.

4.1.2.2 By 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.

Table 4.1 summarises data from the Greater Santiago 1977 Origin Destination Survey (DIC-TUC, 1978); the morning (AM) peak period (the evening peak period is sometimes assumed to be its mirror image) occurred between 7:00 and 9:00 and the representative off-peak period was taken between 10:00 and 12:00. Some comments are in order with respect to this table. Firstly, note that although the vast majority (87.18%) of trips in the AM peak are compulsory (i.e. either to work or education), this is not the case in the off-peak period. Secondly, a typical trait of a developing country emerges from the data: the large proportion of trips for bureaucratic reasons in both periods. Thirdly, a problem caused by faulty classification, or lack of forward thinking at the data-coding stage, is also clearly revealed: the return to home trips (which account for 41.65% of all off-peak trips) are obviously trips with another purpose; the fact that they were returning home is not as important as to why they left home in the first place. In fact, these data needed recoding in order

<table>
<thead>
<tr>
<th>Purpose</th>
<th>AM Peak</th>
<th></th>
<th>Off Peak</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>%</td>
<td>No.</td>
<td>%</td>
</tr>
<tr>
<td>Work</td>
<td>465683</td>
<td>52.12</td>
<td>39787</td>
<td>12.68</td>
</tr>
<tr>
<td>Education</td>
<td>313275</td>
<td>35.06</td>
<td>15567</td>
<td>4.96</td>
</tr>
<tr>
<td>Shopping</td>
<td>13738</td>
<td>1.54</td>
<td>35611</td>
<td>11.35</td>
</tr>
<tr>
<td>Social</td>
<td>7064</td>
<td>0.79</td>
<td>16938</td>
<td>5.40</td>
</tr>
<tr>
<td>Health</td>
<td>14354</td>
<td>1.60</td>
<td>8596</td>
<td>2.74</td>
</tr>
<tr>
<td>Bureaucracy</td>
<td>34735</td>
<td>3.89</td>
<td>57592</td>
<td>18.35</td>
</tr>
<tr>
<td>Accompanying</td>
<td>18702</td>
<td>2.09</td>
<td>6716</td>
<td>2.14</td>
</tr>
<tr>
<td>Other</td>
<td>7736</td>
<td>0.19</td>
<td>2262</td>
<td>0.73</td>
</tr>
<tr>
<td>Return to home</td>
<td>24392</td>
<td>2.72</td>
<td>130689</td>
<td>41.65</td>
</tr>
</tbody>
</table>
to obtain adequate information for trip generation modelling (see Hall et al. 1987). This kind of problem used to occur before the concepts of trip productions and attractions replaced concepts such as origins and destinations, which did not explicitly address the generating capacity of home-based and non-home-based activities.

### 4.1.2.3 By Person Type

This is another important classification, as individual travel behaviour 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 (e.g. 162 in the above example) and this may have *strong implications in terms of data requirements, model calibration and use*. For this reason trade-offs, adjustments and aggregations are usually required (see the discussion in Daly and Ortúzar 1990).

### 4.1.3 Factors Affecting Trip Generation

In trip generation modelling we are typically interested not only in person trips but also in freight trips. For this reason models for four main groups (i.e. personal and freight, trip productions and attractions) are usually required. In what follows we will briefly consider some factors which have been found important in practical studies. We will not discuss freight trip generation modelling, however (although a little had been done by the end of the century), but postpone a discussion on the general topic of freight modelling until Chapter 13.

#### 4.1.3.1 Personal Trip Productions

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 (income, car ownership, household structure and family size) 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; we will come back to this issue in section 4.3.

#### 4.1.3.2 Personal Trip Attractions

The most widely used factor has been roofed space available for industrial, commercial and other services. Another factor used has been zonal employment, and certain studies have attempted to incorporate an accessibility measure. However, it is important to note that in this case not much progress has been reported.

#### 4.1.3.3 Freight Trip Productions and Attractions

These normally account for few vehicular trips; in fact, at most they amount to 20% of all journeys in certain areas of industrialised 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.

### 4.1.4 Growth-factor Modelling

Since the early 1950s several techniques have been proposed to model trip generation. Most methods attempt to predict the number of trips produced (or attracted) by household or zone as a function of (generally linear) relations to be defined from available data. Prior to any comparison of results across areas or time, it is important to be clear about the following aspects mentioned above:
- what trips to be considered (e.g. only vehicle trips and walking trips longer than three blocks);
- what is the minimum age to be included in the analysis (i.e. five years or older).

In what follows we will briefly present a technique which may be applied to predict the future number of journeys by any of the four categories mentioned above. Its basic equation is:

\[ T_i = F_i t_i \]  

(4.1)

where \( T_i \) and \( t_i \) are respectively future and current trips in zone \( i \), and \( F_i \) is a growth factor.
Trip generation modelling

The only problem of the method is the estimation of $F_i$, the rest is trivial. Normally the factor is related to variables such as population ($P$), income ($I$) and car ownership ($C$), in a function such as:

$$F_i = \frac{f(P_i, I_i, C_i)}{f(P_d, I_d, C_d)}$$ (4.2)

where $f$ can even be a direct multiplicative function with no parameters, and the superscripts $d$ and $c$ denote the design and current years respectively.

**Example 4.1:** Consider a zone with 250 households with car and 250 households without car. Assuming we know the average trip generation rates of each group:
- car-owning households produce: 6.0 trips/day
- non-car-owning households produce: 2.5 trips/day

we can easily deduce that the current number of trips per day is:

$$t_i = 250 \times 2.5 + 250 \times 6.0 = 2125 \text{ trips/day}$$

Let us also assume that in the future all households will have a car; therefore, assuming that income and population remain constant (which is a safe hypothesis in the absence of other information), we can estimate a simple multiplicative growth factor as:

$$F_i = C_i^f / C_i^c = 1/0.5 = 2$$

and applying equation (4.1) we can estimate the number of future trips as:

$$T_i = 2 \times 2125 = 4250 \text{ trips/day}$$

However, the method is very crude, as we will demonstrate. If we use our information about average trip rates and make the assumption that these will remain constant (which is actually the main assumption behind one of the most popular forecasting methods, as we will see below), we can estimate the future number of trips as:

$$T_i = 500 \times 6 = 3000$$

which means that the growth factor method would overestimate the total number of trips by approximately 42%. This is very serious because trip generation is the first stage of the modelling process; errors here are carried through the entire process and may invalidate work on subsequent stages.

Growth factor methods are therefore only used in practice to predict the future number of external trips to an area; this is because they are not too many in the first place (so errors cannot be too large) and also because there are no simple ways to predict them. In the following sections we will discuss other (superior) methods which can also be used in principle to model personal and freight trip productions and attractions. However, we will just make explicit reference to the case of personal trip productions as this is the area not only where there is more practical experience, but also where the most interesting findings have been reported.

### 4.2 REGRESSION ANALYSIS

The next subsection provides a brief introduction to linear regression. The reader familiar with this subject can proceed directly to subsection 4.2.2.

#### 4.2.1 The Linear Regression Model

**4.2.1.1 Introduction**

Consider an experiment consisting in observing the values that a certain variable $Y = \{Y_i\}$ takes for different values of another variable $X$. If the experiment is not deterministic we would observe different values of $Y_i$ for the same value of $X_i$.

Let us call $f_i(Y|X)$ the probability distribution of $Y_i$ for a given value $X_i$; thus, in general we could have a different function $f_i$ for each value of $X$ as shown in Figure 4.2. However, such a completely general case is intractable; to make it more manageable certain hypotheses about population regularities are required. Let us assume that:

1. The probability distributions $f_i(Y|X)$ have the same variance $\sigma^2$ for all values of $X$.
2. The means $\mu_i = E(Y|X)$ form a straight line known as the *true regression line* and given by:

$$E(Y_i) = a + bX_i$$ (4.3)

![Figure 4.2] General distributions of $Y$ given $X$
where the population parameters \(a\) and \(b\), defining the line, must be estimated from sample data.

3. The random variables \(Y\) are statistically independent; this means, for example, that a large value of \(Y_1\) does not tend to make \(Y_2\) large.

The above weak set of hypotheses (see for example Wonnacott and Wonnacott 1977) may be written more concisely as:

\[ \text{The random variables } Y_i \text{ are statistically independent with mean } a + bX_i \text{ and variance } \sigma^2. \]

With these Figure 4.2 changes to the distribution shown in Figure 4.3.

\[
E(Y | X) = a + bX
\]

![Figure 4.3 Distribution of Y assumed in linear regression](image)

It is sometimes convenient to describe the deviation of \(Y_i\) from its expected value as the error or disturbance term \(e_i\), so that the model may also be written as:

\[
Y_i = a + bX_i + e_i
\]  
(4.4)

Note that we are not making any assumptions yet about the shape of the distribution of \(Y\) (and \(e\), which is identical except that their means differ) provided it has a finite variance. These will be needed later, however, in order to derive some formal tests for the model. The error term is usual composed of measurement and specification errors (recall the discussion in Chapter 3).

4.2.1.2 Estimation of \(a\) and \(b\)

Figure 4.4 can be labelled the fundamental graph of linear regression. It shows the true (dotted) regression line \(E(Y) = a + bX\), which is of course unknown to the analyst, who must estimate it from sample data about \(Y\) and \(X\). It also shows the estimated regression line \(\hat{Y} = \hat{a} + \hat{b}X\); as is obvious, this line will not coincide with the previous one unless the analyst is extremely lucky (though he will never know it). In general the best he can hope is that the parameter estimates will be close to the target.

\[
\hat{a} = \bar{Y} \quad \text{and} \quad \hat{b} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}
\]  
(4.6)

These estimators have the following interesting properties:

\[
E(\hat{a}) = a \quad \text{Var} (\hat{a}) = \sigma^2/n
\]

\[
E(\hat{b}) = b \quad \text{Var} (\hat{b}) = \sigma^2/\sum x_i^2
\]

In passing, the formula for the variance of \(\hat{b}\) has interesting implications in terms of experimental design. For example, if the \(X\) are too close together, as in Figure 4.5a,
The denominator of (4.9) is usually called the standard error of \( b \) and is denoted by \( s_b \), hence \( t = (b - b_0)/s_b \).

The \( t \)-test  A typical null hypothesis is \( H_0; b = 0 \); in the case (4.8) reduces to:

\[
t = \frac{b - b_0}{s_b}
\]

and this value needs to be compared with the critical value of the Student statistics for a given significance level \( \alpha \) and the appropriate number of degrees of freedom. One problem is that the alternative hypothesis \( H_1 \) may imply unilateral \( b > 0 \) or bilateral \( (b \neq 0) \) tests; this can only be determined examining the phenomenon under study.

Example 4.2: Assume we are interested in studying the effect of income \( I \) in the number of trips by non-car-owning households \( T \), and that we can use the following relation:

\[
T = a + bI
\]

As in theory we can conclude that any influence must be positive (i.e. higher income always means more trips) in this case we should test \( H_0 \) against the unilateral alternative hypothesis \( H_1; b > 0 \). If \( H_0 \) is true, the \( t \)-value from (4.9) is compared with the value \( t_{\alpha/2} \), where \( \alpha \) are the appropriate number of degrees of freedom, and the null hypothesis is rejected if \( t > t_{\alpha/2} \) (see Figure 4.6).

The placement error for the complete model  Figure 4.7a shows the set of values \((\hat{a}, \hat{b})\) for which null hypotheses such as the one discussed above are accepted individually. If we were interested in testing the hypothesis that both estimators are for equal to 0, we
could have a region such as that depicted in Figure 4.7b; i.e., accepting that each parameter is 0 individually does not necessarily mean accepting that both should be 0 together.

Now, to make a two-parameter test it is necessary to know the joint distribution of both estimators. In this case, as their marginals are normal, the joint distribution is also bivariate normal. The $F$-statistic used to test the trivial null hypothesis $H_0: (a, b) = (0, 0)$, provided as one of the standards in commercial computer packages, is given by:

$$F = \left( \frac{\hat{a}^2 + \sum_i x_i^2 \hat{\beta}^2}{2s^2} \right)$$

$H_0$ is accepted if $F$ is less than or equal to the critical value $F_{2, n-2}$. Unfortunately the test is not very powerful (i.e., it is nearly always rejected), but similar ones may be constructed for more interesting null hypotheses such as $(a, b) = (\bar{Y}, 0)$.

### 4.2.1.4 The Coefficient of Determination $R^2$

Figure 4.8 shows the regression line and some of the data points used to estimate it. If no values of $x$ were available, the best prediction of $Y_i$ would be $\bar{Y}$. However, the figure shows that for $x_i$, the error of this method would be high: $(Y_i - \bar{Y})$. When $x_i$ is known, on the other hand, the best prediction for $Y_i$ is $\hat{Y}_i$ and this reduces the error to just $(Y_i - \hat{Y}_i)$, i.e., a large part of the original error has been explained. From Figure 4.8 we have:

$$\begin{align*}
(Y_i - \bar{Y}) &= (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i),
\forall i
\end{align*}$$

Here total deviation = explained deviation + unexplained deviation

If we square the total deviations and sum over all values of $i$, we get the following:

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$

(4.10) total variation explained variation unexplained variation

Now, because $(\hat{Y}_i - \bar{Y}) = \hat{a}x_i$, it is easy to see that the explained variation is a function of the estimated regression coefficient $\hat{b}$. The process of decomposing the total variation into its parts is known as analysis of variance of the regression, or ANOVA (note that variance is just variation divided by degrees of freedom).

The coefficient of determination is defined as the ratio of explained to total variation:

$$R^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}$$

(4.11)

It has limiting values of 1 (perfect explanation) and 0 (no explanation at all); intermediate values may be interpreted as the percentage of the total variation explained by the regression. The index is trivially related to the sample correlation $r$, which measures the degree of association between $X$ and $Y$ (see Wonnacott and Wonnacott 1977).

### 4.2.1.5 Multiple Regression

This is an extension of the above for the case of more explanatory variables and, obviously, more regressors ($\hat{\beta}$ parameters). The solution equations are similar, although more complex, but some extra problems arise which are usually important, such as the following:
1. Multicollinearity. This occurs when there is a linear relation between the explanatory variables; in this case the equations for the regressors $\beta$ are not independent and cannot be solved uniquely.

2. How many regressors to include. To make a decision in this case, several factors have to be taken into consideration:
   - Are there strong theoretical reasons to include a given variable, or is it important for policy testing with the model?
   - Is the variable significant (i.e. $H_0$ rejected in the t-test) and is the estimated sign of the coefficient consistent with theory or intuition?

If in doubt, one way forward is to take out the variable and re-estimate the regression in order to see the effect of its removal on the rest of the coefficients; if this is not too important the variable can be left out for parsimony (the model is simpler and the rest of the parameters can be estimated more accurately). Commercial software packages provide an ‘automatic’ procedure for tackling this issue (the stepwise approach); however, this may induce some problems, as we will comment below. We will come back to this general problem in section 8.3 (Table 8.1) when discussing discrete choice model specification issues.

3. Coefficient of determination. This has the same form as (4.11). However, in this case the inclusion of another regressor always increases $R^2$; to eliminate this problem the corrected $R^2$ is defined as:

$$R^2 = \frac{|R^2 - k/(n-1)|}{(n-1)/(n-k-1)}$$

(4.12)

where $\hat{n}$ stands for sample size as before and $k$ is the number of regressors $\hat{\beta}$.

In trip generation modelling the multiple regression method has been used both with aggregate (zonal) and disaggregate (household and personal) data. The first approach has been practically abandoned in the case of trip productions, but it is still the premier method for modelling trip attractions.

4.2.2 Zonal-based Multiple Regression

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 behaviour 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 had an homogeneous socioeconomic composition, but represented as wide as possible a range of conditions. A major problem is that the main variations in person trip data occur at the intra-zonal level.

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 HB 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:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i$$

whereas the model using rates would be:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i$$

with $y_i = Y_i/H_i; x_i = X_i/H_i; \epsilon_i = E_i/H_i$ and $H_i$ the number of households in zone $i$.

Both equations are identical, in the sense that they seek to explain the variability of trip making behaviour 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 $H_i$ 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 $1/H_i$, 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 (see Douglas and Lewis 1970). What is certainly unsound is the mixture of means and aggregate variables in a single model.

To end this theme it is important to remark that even when rates are used, zonal regression is conditioned by the nature and size of zones (i.e. the spatial aggregation problem). This is clearly exemplified by the fact that inter-zonal variability diminishes with zone size as shown in Table 4.2, constructed with data from Perth (Douglas and Lewis 1970).
Trip generation modelling

Table 4.2  Inter-zonal variation of personal productions for two different zoning systems

<table>
<thead>
<tr>
<th>Zoning system</th>
<th>Mean value of trips/household/zone</th>
<th>Inte-r-zonal variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>75 small zones</td>
<td>8.13</td>
<td>5.85</td>
</tr>
<tr>
<td>23 large zones</td>
<td>7.96</td>
<td>1.11</td>
</tr>
</tbody>
</table>

4.2.3 Household-based Regression

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 believed 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). We will challenge this thesis in section 4.3.3.

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 behaviour. 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 4.3: Consider the variables trips per household (Y), number of workers (X1) and number of cars (X2). Table 4.3 presents the results of successive steps of a step-wise model estimation, the last row also shows (in parenthesis) values for the t-ratio (equation 4.9). Assuming 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 Y should increase with both X1 and X2).

The third model is a good equation in spite of its low R². The intercept 0.91 is not large (compare it with 1.44 times the number of workers, for example) and the regression coefficients are significantly different from zero (H0 is rejected in all cases). The model could probably benefit from the inclusion of other variables.

An indication of how good these models are may be obtained from comparing observed and modelled trips for some groupings of the data (see Table 4.4). 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 bias 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.

Table 4.4  Comparison of trips per household (observed/estimated).

<table>
<thead>
<tr>
<th>No. of cars</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3 or more</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.9/0.9</td>
<td>2.1/2.4</td>
<td>3.4/3.8</td>
<td>5.3/5.6</td>
</tr>
<tr>
<td>1</td>
<td>3.2/2.0</td>
<td>3.5/3.4</td>
<td>3.7/4.9</td>
<td>8.5/6.7</td>
</tr>
<tr>
<td>2 or more</td>
<td></td>
<td>4.1/4.6</td>
<td>4.7/6.0</td>
<td>8.5/7.8</td>
</tr>
</tbody>
</table>

4.2.4 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 4.9 presents data for households stratified by car ownership and number of workers. It can be seen that travel behaviour is non-linear with respect to family size.

It is important to mention that there is a class of variables, those of a qualitative nature, which usually shows non-linear behaviour (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 linearise their effect (e.g. take logarithms, raise to a power). However, selecting the most adequate transformation is not an easy or arbitrary exercise, 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 behaviour. 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 4.4: Consider the model of Example 4.3 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 4.3 would now be:

\[
Y = 0.84 + 1.41X_1 + 0.75Z_1 + 3.14Z_2 \\
\begin{array}{ccc}
(3.6) & (8.1) & (3.2)
\end{array}
\]

Even without the better \(R^2\) value, this model is preferable to the previous one just because the non-linear 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 4.10.

Looking at Figure 4.10, the following question arises: would it not be preferable to estimate separate regressions for the data on each group as in that case we would not require each line to have the same slope (the coefficient of \(X_1\))? The answer is in general no unless we had a reasonable amount of data for each class. The fact is that the model with dummies uses all the data, while each separate regression would use only part of it, and this is in general disadvantageous. It is also interesting to mention that the use of dummy variables tends to reduce problems of multicollinearity in the data (see Douglas and Lewis 1971).

4.2.5 Obtaining Zonal Totals

In the case of zonal-based regression models, this is not a problem as the model is estimated precisely at this level. In the case of household-based models, though, an aggregation stage is required. Nevertheless, precisely because the model is linear the aggregation problem is trivially solved by replacing the average zonal values of each independent variable in the model equation and then multiplying it by the number of households in each zone. However, it must be noted that the aggregation stage can be a very complex matter in non-linear models, as we will see in Chapter 9.

Thus, for the third model of Table 4.3 we would have:

\[
T_i = H_i(0.91 + 1.44\bar{X}_{1i} + 1.07\bar{X}_{2i})
\]

where \(T_i\) is the total number of HB trips in the zone, \(H_i\) is the total number of households in it and \(\bar{X}_{ji}\) is the average value of variable \(X_j\) for the zone.

On the other hand, when dummy variables are used, it is also necessary to know the number of households in each class for each zone; for instance, in the model of Example 4.4 we require:

\[
T_i = H_i(0.84 + 1.41\bar{X}_{1i}) + 0.75H_{1i} + 3.14H_{2i}
\]

where \(H_j\) is the number of households of class \(j\) in zone \(i\).

This last expression allows us to appreciate another advantage of the use of dummy variables over separate regressions. To aggregate the models, in this latter case,
it would be necessary to estimate the average number of workers per household \( (X_t) \) for each car-ownership group in each zone, and this may be complicated.

### 4.2.6 Matching Generations and Attractions

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

\[
\sum_i O_i = \sum_j D_j \quad (4.13)
\]

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 (see Chapter 5).

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 is 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 \( O_t \) is in fact the correct figure for \( T \); therefore, all destinations \( D_j \) are multiplied by a factor \( f \) given by:

\[
f = \frac{T}{\sum_j D_j} \quad (4.14)
\]

which obviously ensure that their sum adds to \( T \).

### 4.3 CROSS-CLASSIFICATION OR CATEGORY ANALYSIS

#### 4.3.1 The Classical Model

##### 4.3.1.1 Introduction

Up to the late 1960s most transportation planning studies in the USA developed trip generation equations based on linear regression analysis, particularly when modelling personal trip productions. In fact, the regression model was favoured as the central method in the Federal Highway Administration's guide to trip generation analysis (FHWA 1967).

At the end of the 1960s an alternative method for modelling trip generation appeared and quickly became established as the preferred one in the United Kingdom. The method was known as category analysis in the UK (Wootton and Pick 1967) and cross-classification in the USA; there it went through a similar development process as the linear regression model, with earliest procedures being at the zonal level and subsequent models based on household information.

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. Although the method was originally designed to use census data in the UK, a serious problem of the approach remains the need to forecast the number of households in each strata in the future.

##### 4.3.1.2 Variable Definition and Model Specification

Let \( \Phi(h) \) 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 \( T(h) \) by purpose group. The rate \( \Phi(h) \) is then the total number of trips in cell \( h \), by purpose, divided by the number of households \( H(h) \) in it.

In mathematical form it is simply as follows:

\[
\Phi(h) = \frac{T(h)}{H(h)} \quad (4.15)
\]

The 'art' of the method lies in choosing the categories such that the standard deviations of the frequency distributions depicted in Figure 4.11 are minimised.

![Figure 4.11 Trip-rate distribution for household type](image)
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 monotonic, 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. There are no statistical goodness-of-fit measures for the model, so only aggregate closeness to the calibration data can be ascertained.
3. 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. For example, in the Monmouthshire Land Use/Transportation Study (see Douglas and Lewis 1971) the following distribution for 108 categories (six income levels, three car ownership levels and six household structure levels) was found, using a sample of 4000 households:

<table>
<thead>
<tr>
<th>Table 4.5 Household frequency distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of categories</td>
</tr>
<tr>
<td>No. of households surveyed</td>
</tr>
</tbody>
</table>

Accepted wisdom suggests that at least 50 observations per cell are required to estimate the mean reliably; thus, this criterion would be satisfied in only 18 of the 108 cells for a sample of 4000 households. There may be some scope for using stratified sampling to guarantee more evenly distributed sample sizes in each category. This involves, however, additional survey costs.

There is no effective way to choose among variables for classification, or to choose best groupings of a given variable; the minimisation of standard deviations hinted at in Figure 4.11 requires an extensive ‘trial and error’ procedure which may be considered infeasible in practical studies.

4.3.13 Model Application at Aggregate Level

Let us denote by \( n \) the person type (i.e. with and without a car), by \( a_i(h) \) the number of households of type \( h \) in zone \( i \), and by \( H^p(h) \) the set of households of type \( h \) containing persons of type \( n \). With this we can write the trip production with purpose \( p \) by person type \( n \) in zone \( i \), \( O^p_{i,h} \), as follows:

\[
O^p_{i,h} = \sum_{h \in H^p(h)} a_i(h) \phi(h)
\]  

(4.16)

To verify how the model works it is possible to compare these modelled values with observed values from the calibration sample. Inevitable errors are due to the use of averages for the \( \phi(h) \); one would expect a better stratification (in the sense of minimising the standard deviation in Figure 4.11) to produce smaller errors.

There are various ways of defining household categories. The first application in the UK (Wootton and Pick 1967) 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 4.6:

<table>
<thead>
<tr>
<th>Table 4.6 Example of household structure grouping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
</tbody>
</table>

The problem is clearly how to predict the number of households in each category in the future. The method most commonly used (see Wilson 1974) consists in, firstly, defining and fitting to the calibration data, probability distributions for income \( I \), car ownership \( C \) and household structure \( S \); secondly, using these to build a joint probability function of belonging to household type \( h = (I, C, S) \). Thus, if the joint distribution function is denoted by \( \phi(h) = \phi(I, C, S) \), the number of households in zone \( i \) belonging to class \( h, a_i(h) \), is simply given by:

\[
a_i(h) = H_i \phi(h)
\]  

(4.17)

where \( H_i \) is the total number of households in the zone. This household estimation model may be partially tested by running it with the base-year data used in calibration. The total trips estimated with equation (4.16), but with simulated values for \( a_i(h) \), can then be checked against the actual observations.

One further disadvantage of the method can be added at this stage:

5. If it is required to increase the number of stratifying variables, it might be necessary to increase the sample enormously. For example, if another variable was added to the original application discussed above and this was divided into three levels, the number of categories would increase from 108 to 324 (and recall the discussion on Table 4.5).
4.3.2 Improvements to the Basic Model

4.3.2.1 Multiple Classification Analysis (MCA)

MCA is an alternative method for defining classes and testing 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. The interested reader is referred to Stopher and McDonald (1983) for full details as only a summary is provided below.

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 terms 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 4.5: Table 4.7 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, columns the grand average.

Table 4.7 Number of households per cell and mean trip rates

<table>
<thead>
<tr>
<th>Household size</th>
<th>0 car</th>
<th>1 car</th>
<th>2+ cars</th>
<th>Total</th>
<th>Mean trip rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 person</td>
<td>28</td>
<td>21</td>
<td>0</td>
<td>49</td>
<td>0.47 (±0.07)</td>
</tr>
<tr>
<td>2 or 3 persons</td>
<td>150</td>
<td>201</td>
<td>93</td>
<td>444</td>
<td>1.28 (±0.18)</td>
</tr>
<tr>
<td>4 persons</td>
<td>61</td>
<td>30</td>
<td>75</td>
<td>266</td>
<td>1.86 (±0.34)</td>
</tr>
<tr>
<td>5 persons</td>
<td>37</td>
<td>142</td>
<td>90</td>
<td>269</td>
<td>1.90 (±0.36)</td>
</tr>
<tr>
<td>Total</td>
<td>276</td>
<td>454</td>
<td>258</td>
<td>988</td>
<td>1.54</td>
</tr>
</tbody>
</table>

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; 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 4.8 depicts the full trip-rate table together with its row and column deviations.

Table 4.8 Trip rates calculated by multiple classification

<table>
<thead>
<tr>
<th>Car ownership level</th>
<th>0 car</th>
<th>1 car</th>
<th>2+ cars</th>
<th>Deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>HM size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 person</td>
<td>0.00</td>
<td>0.46</td>
<td>1.37</td>
<td>–1.07</td>
</tr>
<tr>
<td>2 or 3 persons</td>
<td>0.48</td>
<td>2.17</td>
<td>2.18</td>
<td>–0.26</td>
</tr>
<tr>
<td>4 persons</td>
<td>1.05</td>
<td>1.85</td>
<td>2.76</td>
<td>0.32</td>
</tr>
<tr>
<td>5 persons</td>
<td>1.09</td>
<td>1.89</td>
<td>2.80</td>
<td>–0.56</td>
</tr>
<tr>
<td>Deviations</td>
<td>–0.81</td>
<td>–0.01</td>
<td>0.90</td>
<td></td>
</tr>
</tbody>
</table>

In contrast 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.

The most important statistical goodness-of-fit measures associated with MCA are (Stepher and McDonald 1983):

- an $F$-statistic to assess the entire classification scheme (recall the discussion in section 4.2);
- a correlation ratio statistic for assessing the contribution of each classification variable (see Stepher 1975); and
- an $R^2$ measure for the complete cross-classification model.

These measures allow the analyst to compare different classification schemes and to assess their fit to the base-year data.

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 4.6: Table 4.9 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 4.8.

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 4.9 (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 4.8. Note that they could have arisen by problems of small sample size at least in one case.

Table 4.9  Trip rates for same study area calculated using category analysis

<table>
<thead>
<tr>
<th>HH size</th>
<th>Car ownership level</th>
<th>0 car</th>
<th>1 car</th>
<th>2+ cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 person</td>
<td></td>
<td>0.12</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>2 or 3 persons</td>
<td></td>
<td>0.00</td>
<td>1.28</td>
<td>2.16</td>
</tr>
<tr>
<td>4 persons</td>
<td></td>
<td>1.14</td>
<td>1.74</td>
<td>2.60</td>
</tr>
<tr>
<td>5 persons</td>
<td></td>
<td>1.02</td>
<td>1.69</td>
<td>2.60</td>
</tr>
</tbody>
</table>

If there are interactions between the stratifying variables, an adjusted version of the MCA approach can be used (Stopher and McDonald 1983). This simply consists of calculating the mean trip rate for each level of car ownership and household size (or income if the stratification considers it), weighting by the proportion of households (or individuals) at that level (independently of the levels of the remaining stratifying variables), and then using the normal MCA computations to estimate the trip rates per cell. This allows interactions to be taken into account, because the weighted averages decrease the size of the deviation with respect to the general mean. Ortizar et al. (1998) used this method in a study of the stability of trip generation rates in Santiago, Chile, for classes formed by the interaction of income (five levels), car ownership (three levels) and family size (three levels); they found that it outperformed the normal MCA.

4.3.2.2 Regression Analysis for Household Strata

A mixture of cross-classification and regression modelling of trip generation may be the most appropriate approach on certain occasions. For example, in an area where the distribution of income is unequal it may be important to measure the differential impact of policies on different income groups; therefore it may be necessary to model travel demand for each income group separately throughout the entire modelling process. Assume now that in the same area car ownership is increasing fast and, as usual, it is not clear how correlated these two variables are; a useful way out may be to postulate regression models based on variables describing the size and make-up of different households, for a stratification according to the two previous variables.

Example 4.7: Table 4.10 presents the 13 income and car-ownership categories (C) defined in ESTRATUS (1989) for the Greater Santiago 1977 origin-destination data. As can be seen, the bulk of the data corresponds to households with no cars and low income. Also note that categories 7 and 10 have rather few data points; this is, unfortunately, a general problem of this approach. Even smaller samples for very low income and high car ownership led to the aggregation of some categories at this range.

The independent variables available for analysis (i.e., after leaving out the stratifying variables) included variables of the stage in the family cycle variety, which we will discuss in section 4.4. However, after extensive specification searches it was found that the most significant variables were: number of workers (divided into four classes depending on earnings and type of job), number of students and number of residents.

Linear regression models estimated with these variables for each of the 13 categories were judged satisfactory on the basis of correct signs, small intercepts, reasonable significance levels and R² values (e.g., between 0.401 for category 4, and 0.682 for category 7; see Hall et al. 1987).

4.3.3 The Person-category Approach

4.3.3.1 Introduction

This is an interesting alternative to the household-based models discussed above, which was originally proposed by Supernak (1979). It has been argued that this approach offers the following advantages (Supernak et al. 1983):

1. A person-level trip generation model is compatible with other components of the classical transport demand modelling system, which is based on tripmakers 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.
3. The sample size required to develop a person-category model can be several times smaller than that required to estimate a household-category model.
4. 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.
5. Person categories are easier to forecast than household categories as the latter require forecasts about household formation and family size; these tasks are