such variables are required and how they appear in logit models. The paper also shows the technical problems that can arise in estimating models containing multiple size variables and the algorithm that has been adopted to deal with these problems. Finally, the paper gives some practical evidence (based on data from another study) that suggests that allowing for the proper estimation of size variables can be of significant advantage in practice, compared with previous methodology.

The evidence from the paper cited and the ready availability of the software were reasons to use the software for the Zuidvleugel study. Incorporation of size variables in the models could therefore be based solely on theoretical and empirical considerations, rather than being constrained by the technical capabilities of the estimation software, as had been the case in past studies. Several of the models in fact contain multiple attraction variables.

4.4 SAMPLING

Sampling is an important technique in almost every statistical study. In most cases where the studied population is large, the cost of obtaining information from the whole population is prohibitive and it is fortunate that there exist statistical techniques that allow analysis with a much smaller data set. In every case, the issue arises of what sample size is appropriate and this turns on the cost of increasing the sample size set aqainst the inaccuracy of reducing the sample. Any possibility of obtaining more information (more accuracy) by choosing more carefully a sample of a given size is clearly worth investigation.

For the Zuidvleugel study, the most fundamental sample was the 2949-household data set selected for the home interview. The size of this data set had been determined by considerations of cost and modelling experience, but (as described in Report 4) the selection of the actual households was a three-stage stratified random sample already designed to focus the study resources on the most relevant travellers. To reduce costs further, however, further sampling was undertaken within this data set, and attempts were again made to focus on the most relevant travellers. A second type of sampling was also undertaken for the same reason, which was sampling of the alternatives available to travellers that is, when one zone had been selected from the 396 available, a small number (20) of the unselected zones were included in the analysis. Again attempts were made to maximise the efficiency of this sampling. In the two following sections these two sampling processes are described and explained, with reference to the

theoretical literature and the practical problems of this particular study.

4.4.1 SAMPLING OBSERVATIONS

Further sampling of observations within the home interview data set raised a particular problem for the mode choice models. This problem was that public transport took a rather small share and any reduction in the number of public transport tours was therefore undersirable. But although sampling stratified by exogenous variables (i.e. not directly dependent on the choice of mode) raises no special problems for estimation, sampling stratified by choice, as was necessary in this case, does raise special problems. These problems have been studied by Manski and Lerman*.

This important paper established the propriety and practicality of using choice-based samples by introducing an estimator and proving its consistency with choice-based data. The estimator works by introducing weights into the estimation, which would have been possible in the Zuidvleugel study. The basic Manski-Lerman result is applicable to all choice probability models. An additional result given in that paper (attibuted to McFadden) is that if the model is of logit form and if it has a full set of alternative-specific constants, then the only biases introduced by using the simple (unweighted) estimator are in those constants and the constants can be corrected rather simply.

For the Zuidvleugel study, this latter alternative was more attractive and it was therefore adopted. This alternative was preferable because the absence of weights allowed the existing software to be used unchanged.

4.4.2 SAMPLING ALTERNATIVES

The Zuidvleugel study made, as noted previously, geographical distinctions for representing destinations, defining 396 zones for the Netherlands. This large number was a particular difficulty for model estimation, since data has to be held on the computer for every alternative. In the case of joint mode-destination choice model with three modes, (say) five hundred observations and twenty variables, a data file with nearly 12 million words would have to be stored and processed. It was therefore most desirable to reduce

---

* C.F. Manski and S.R. Lerman, The Estimation of Choice Probabilities from Choice Bases Samples, Econometrica Vol 5 No 8 (November 1977).

the burden of computation.

A procedure that has been used in a number of previous studies is to sample the alternatives. This is a natural step if the logit function is considered, giving the probability of the observed choice as

$$\log P_c = V_c - \log \sum_{K \varepsilon S} \exp V_k \qquad (9)$$

where is S the choice set and V_k is the measured part of the utility of choice k. The logsum variable can be estimated by drawing a sample set T from the choice set S calculating the sum of exponentials over the sample set and factoring up to represent the whole set. It is natural that when S becomes larger then T should also become larger to maintain the same level of accuracy. Little is known about how the size of T should be chosen to achieve a specified accuracy of estimation.

An important point that is relevant for the sampling strategy is the large physical size of the Zuidvleugel area. This large size, together with the fact that the mean tour length is comparatively short, means that adding zones on the far side of the study area adds little to the accuracy of the estimate. In terms of the logit equation (9), there is little point in collecting information about many of the smaller values of exp V_k because they add so little to the total sum. A sample was therefore designed in which the more "important" alternatives are sampled with higher probability than the less "important".

The "importance" of an alternative k may reasonably be represented by exp V_k, which is proportional to the probability of choice as given by the model but which is not known. The sampling is based on an approximate preliminary estimate, but the exact optimisation of the sampling is not important to the overall efficiency of the process. The approximate preliminary models in this study used only distance and a measure of the "size" of each zone. Three sampling schemes were in fact employed, using distance and

- zonal employment, for work-related purposes;
- recreational area, for recreation tours;
- population, for other purposes.

Different parameters were used for the reduction of probability with distance in each case, to reflect the different tour lengths in each group.

In each case it was decided to sample 20 zones. This number was determined primarily as being the largest number giving computer files of acceptable size (about 0.5 million words). Although these computer files gave some diffi-

culties in practical processing, the lack of knowledge about the modelling implications of different sample sizes made it advisable to use the largest possible sample.

Drawing a sample for each of the 11000 tours being modelled would again have given too large computer files. Accordingly, samples were drawn (for each sample scheme) for each of the 730 location at which households were represented. According to the theoretical basis of sampling alternatives, such a restriction on the sampling will introduce no biases.

To check the adequacy of the samples, given that there is no theory indicating appropriate sizes, a number of tests were made. These showed that the frequency with which the chosen alternative appeared in the sample ranged from about 70% for work purposes to over 90% for other purposes. Similar percentages were obtained when the sum of exp V_k calculated over the sample set was compared with the sum calculated over the total set. These percentages contrast with the 5% that would be obtained by uniform sampling and suggest that the great (and necessary) reduction is computer processing obtained by sampling has not caused too great a loss in the information input to the modelling.

During the estimation process it is necessary to adjust the likelihood function to take account of the importance sampling. In this study this was done by factoring exp V_k for each destination k, by the inverse of the probability with which the destination was selected as calculated from the importance sampling. The best theoretical work available on the subject* suggests a different factoring, in which the factoring is based on the entire set selected. Relative to the factors given in the theory, those used in this study represent approximations. Because the approximation to the logsum variable is in any case very good (as explained above), it is not likely that important defects have been introduced by the use of the much simpler factors used. Calculation of theoretically correct factors would be a challenging problem for a model system of this scale.

4.5 AGGREGATION

All of the models described in this report are disaggregate, in the sense that

---

* D. McFadden, Modelling the Choice of Residential Location, in Karlqvist et al (eds) Spatial Interaction Theory and Planning Models, North Holland, Amsterdam (1978).

the behaviour represented is that of a specific individual. These models indicate the most important influences on the behaviour of individuals, but most applications require forecasts for entire populations. The present section describes the methods used to obtain population forecasts from the disaggregate models.

Two methods were used, each having specific advantages for specific policy applications. The first method, based on zones, closely parallels the traditional zonal modelling often used in transportation planning. For the Zuidvleugel Study, a more extensive market segmentation was used than is usual, in an attempt to improve the accuracy of the models. This zonal aggregation is particularly suitable for policy issues relating strongly to specific locations, and especially where assignments are required. The second method, called "sample enumeration", is based on the application of the models to a sample of the population being modelled, and the expansion of that sample to represent the entire population. The sample enumeration method is particularly suitable when more general policy issues (e.g. pricing) are under consideration.

4.5.1 ZONAL MODELS

The objective of zonal models is to produce a travel matrix: that is, a matrix giving for each mode, for every pair of zones defined, the amount of travel by that mode originating in the first zone with destination in the second zone. For the Zuidvleugel Study a total of 402 zones were defined, of which 396 were in the Netherlands and 319 in the specific Zuidvleugel area. The objective of the zonal models is thus to predict the number of tours orginating in each of the 319 study area zones with destination in each of the 396 zones of the Netherlands (cross-border tours were added subsequently).

The method applied for zonal modelling is to construct for each zone a "typical traveller" ,to use the models to forecast the behaviour of this typical traveller (i.e. frequency, destination and mode choice) and to expand by the number of travellers in the zone. For increased accuracy a number of market segments, repesenting different car availability groups, etc., may be defined and a typical traveller modelled in each group. For the Zuidvleugel study a total of 20 market segments were defined, allowing representation of variations with car availability, age and sex.

There are two limitations on the application of zonal models. The first is

the large number of calculations that must be made. For example, with 20 market segments and 400 zones, the mode choice model must be applied 3.2 million times for each travel purpose. This huge amount of calculation clearly limits the accuracy with which "typical travellers" can be defined in terms of socio-economic or locational characteristics, since any increase in accuracy would increase the number of segments or the number of zones. The second limitation is that even if it were possible to perform the required calculations, detailed socio-economic data is simply not available at a zonal level or is very expensive to collect. These two limitations mean that forecasting at a zonal level must inevitably be based on coarser information than is desirable. This has consequences for the accuracy of the zonal models, but also for the way in which they are constructed.

From a household interview it is relatively easy to collect much more socio-economic data than can be represented in a zonal model. Moreover, the accuracy with which the location of the household is known is much better than the rather large zones that are inevitable in a regional study. Given this wealth of detailed information, it is possible to construct much better models than those that can be derived from the data available at zonal level. These fully disaggregate models are not only better in that they include important effects that cannot be represented at zonal level (e.g. competition within the household for cars), but also that even the effects that <u>can</u> be represented at zonal level are more accurately measured when all the available disaggregate information is taken into the model.

Thus even when it is required only to produce models at the zonal level, it is an useful procedure first to estimate fully disaggregate models, then to fine some means of aggregation, than to work with data defined at a zonal level.

In practice, the variables used in the fully disaggregate models fall into two classes: those for which data are available at zonal level (probably not so accurately as at the disaggregate level) and those for which no data is available at zonal level. In principle, there is no difference even between these classes: for every variable in the fully disaggregate model a zonal equivalent must be found; this equivalent is more or less accurate depending on the circumstances of the particular variable (maybe only an overall average). In each case there is a loss of accuracy which ranges from none to total: in short, it is a question of degree.

It is essential to note that in changing the model from operating with fully disaggregate to zonal variables, there must be adjustments to the model to compensate and correct for the loss of accuracy. These changes are necessary for two reasons:

- the zonal variables may have different average values from the fully disaggregate variables, thus requiring a change in the alternative-specific constants of the model;
- the zonal variables will have, as noted above, a large or small loss in accuracy relative to the fully disaggregate variables. A failure to correct the scale of the utility functions in the model will result in the overprediction of response to changes known as "aggregation bias".

Aggregation bias arises essentially because the model is non-linear, and with a non-linear model a prediction for an average value is not the same as the average of the model's predictions for individual values. In the case of S-shaped response curves such as the logit model this typical aggregation problem has the important effect of causing a consistent over-prediction of response to changes. This effect is naturally very undesirable and a correction to the scale of the utility functions is therefore essential.

From a theoretical standpoint, the best procedure for constructing a zonal model would be:

1. Estimate the best model using all the fully disaggregate data available.
2. Using the coefficients estimated in step 1 and the best available zonal data, construct an approximate utility for each alternative.
3. Using as variables only the utilities constructed in step 2, estimate a new choice model, giving a scaling parameter for all the utilities and new constants for the alternatives.

This procedure takes advantage in step 1 of all the available data to calculate the best possible coefficients for all variables. These are then corrected by a single scaling factor estimated in step 3, which theoretically must be less than 1 (but positive), which eliminates the aggregation bias. This standard procedure was used to derive the zonal models from fully disaggregate models in a number of cases, but in other cases it was preferable to adopt a slightly different approach.

The need for a different approach to aggregation in some cases comes from the very limited data available at zonal level. For a number of socio-economic variables, no approximation to the variable used in the fully disaggregate

model was available at zonal level, although another variable might be available which gave some of the information in a totally different form. For example, a fully disaggregate model might contain an income variable: at the zonal level the only variable available was an indication of whether the zonal income was on average low, middle or high. The inclusion of the zonal indicator variable can improve the quality of the zonal model, but it is not clear how this variable should be related to the fully disaggregate variable. This type of difficulty arose frequently with the socio-economic variables.

For this reason, in a number of cases, zonal models were simply estimated without formal input from the fully disaggregate models. This method is in practice rather simpler than the theoretically prefered procedure outlined above. For the socio-economic variables, a simple estimation at zonal level allows the limited data to be used to full effect. For the level-of-service variables, the coefficients obtained need to be checked against those obtained from the fully disaggregate model to ensure that the values are consistent with theoretical expectation: the zonal values should be a constant multiple (between zero and one) of the fully disaggregate values. In most cases this check was satisfactory.

4.5.2 SAMPLE ENUMERATION

Sample enumeration is an approach to deriving population forecasts from disaggregate travel demand models. The technique consists of the application of the models to a representative sample of travellers, and the expansion of that sample to give estimates of population totals.

The chief advantage of sample enumeration from a theoretical point of view is that there is no strong constraint on the amount of information that can be considered for each individual. Fully disaggregate models can therefore be applied. This characteristic avoids the need for an aggregation step (as explained above for zonal models). Further, it allows the use in forecasting of the best models available.

From a practical standpoint, sample enumeration is attractive in that it allows the analyst to control the relative cost and accuracy of his forecast. Working with a sample of a few hundred individuals will normally be satisfactory, and the computer time required for such a run is a small fraction of that required for a zonal model application. By choosing the sample size, the time and cost of an application can be balanced against the accuracy needed for a particular

application.

The disadvantage of sample enumeration is that a representative sample must be available allowing sufficiently accurate forecasts to be made for the policy under test. For short-term policies applying to the study area as a whole, the home interview survey used for model estimation forms an ideal basis for sample enumeration forecasting. For longer-term policies or for policies in which specific locations are important - particularly for assignments - finding a suitable sample is more difficult.

For longer-term policies, what is required is to construct an artifical sample that is believed to be representative for the forecast year. Some work on the contruction of such samples has been done in the United States and in Britain and further work is in hand and planned in the Netherlands. Sample enumeration appears particularly suitable for long-term analysis, where very specific areas are in any case difficult to study. It is likely that it will soon be possible to apply the technique routinely for a future year as it is now for current years.

Application of sample enumeration as a basis for assignments seems even in principle less likely to become practical. An assignment requires so much detail of location that the sample size necessary to achieve reasonable accuracy would be very large. It is conceivable that for an assignment across a large natural barrier with few crossings - e.g. the Hollands Diep - a sample enumeration could be used to predict flows on the crossings only. But for area-wide assignments a zonal model remains the best available technique for practical studies.

4.6 TESTS OF MODEL SPECIFICATIONS

Development of travel demand models involves both theoretical and empirical considerations. Report 1 and the preceding sections of this report describe the theoretical considerations preliminary to estimation of the models. Report 5 gives some empirical findings from the home interview survey that led to decisions on model structure. These two aspects must be considered simultaneously when models are actually being estimated.

Ultimately, a large amount of judgement enters into the selection of the final model of each decision. On the one hand, the models have been developed with certain concepts of behaviour, and some of these are so firmly rooted

that a model in conflict with them would be rejected on theoretical grounds. For example, although the models are not entirely dependent on the notion of individual utility maximisation, they are expected to be sufficiently consistent with it that a model suggesting positive marginal utilities for travel times or costs would be rejected. On the other hand, the initial theory must not be so rigid as to deny the possibility of learning from the data collected. The models finally selected must therefore embody a theory balanced between the essential a priori axioms and the new insight gained from the data.

Although, as noted, there is an important element of professional judgement in the selection of the final models, there are also a number of more or less formal tests can be made explicitly to test models during development. These tests can indicate suspicious data items, model specification errors, directions for improvement and the acceptability of the model as a whole, or can be used to compare one model with another.

4.6.1 INFORMAL TESTS

The first test made on any estimated model is a check on the values of the coefficients. As mentioned, theoretical arguments will imply signs for a number of the coefficients, and a sign contrary to expectations indicates an error in the theoretical argument or some specification error in the model. All such inconsistent signs were eliminated from the models.

A second check that can be made on the coefficients is to examine their relative value. For example, the relative values of time and cost coefficients should be acceptable. The models should not be considered as means of deriving values of time, since the data has not been collected,nor the models developed,with this in mind, and the accuracy is correspondingly not ideal. The values implied should nevertheless fall within a reasonable range. Similarly, the coefficients of socio-economic variables should be reasonable relative to each other. A failure of a model to meet this test will usually indicate a specification error.

A further test on the coefficients is to examine the elasticities implied by the model. This tests the absolute values of the coefficients.

A test that is particularly important in logit modelling is to investigate the existence of "outliners". These are observations of travellers whose

choices seems particularly unlikely, often as a result of coding errors in the data. It can be shown that these observations have a disproportionately large effect on the values of the coefficients* so that it is essential to check in each model that coding errors are not excessively disturbing the estimates. In this Study, a number of observations were eliminated for this reason.

A further test, which proved extremely useful in indicating specification errors, was to tabulate the choice proportions predicted by the models together with the observed choice proportions for each of a series of classifications of the data (e.g. by income or car availability). This type of disaggregate validation of the model is particularly useful in indicating directions for improving the specification. Extensive tabulations were made for every model estimated.

4.6.2 FORMAL TESTS

The coefficients of the models are estimated by the standard maximum likelihood method. This gives not only estimates of the values of the coefficients, but also good information about the accuracy of those estimates. Further, inferences may be drawn from the values of the likelihood function itself.

The direct check on the accuaracy of estimation is the standard error of each estimated value. This can be used in the usual way to give confidence limits within which the true value of the coefficient is believed to lie (using the fact the distribution of the estimates is approximately multivariate normal in the vicinity of the optimum).

A particularly interesting question for many variables is whether the value zero is within acceptable confidence limits. The acceptability of zero as a value can help to determine whether or not a variable should be included in the model, for if zero is not within acceptable confidence limits, the hypothesis that the true value of the coefficient is zero (i.e. that the variable should be omitted) can be rejected. A useful means of expressing the acceptability of zero as a value is to construct the (approximate) "t-ratio" being the value of the coefficient divided by its standard error, i.e. the number of standard errors between the best estimate and zero (the t-ratio is approx-

---

* A.J. Daly and S. Zachary, "Commuters' Values of Time" Report Local Government O.R. Unit, Reading, England 1975.

imate because (unlike linear regression) the distribution of the estimates is only approximately normal). These t-ratios have been calculated for every variable for which the issue of inclusion or not inclusion in the model is important, and are reported in Chapter 5.

In the case of logsum variables, it is also of interest whether the coefficient value is significantly different from 1. If the value 1 is not acceptable, then the data reject the alternative structure in which upper and lower level choices are represented jointly (in the same logit model). The notation t_1-ratio has been coined to express the number of standard deviations between the best estimate value and 1, and this ratio can be used in the same way as the normal t-ratio (sometimes called t_0-ratio for clarity). The t_1-ratios have been calculated for every logsum variable, and are reported in Chapter 5.

For some variables, such as alternative-specific constants, the value zero has no special significance. This is because, in the case of constants, it is known a priori that many important alternative-specific effects have been omitted from the model. For the net balance of these effects to be found to be zero would be a numerical coincidence without importance for explaining behaviour or for determining an appropriate model structure. The test as to whether zero is acceptable is not therefore interesting and t-ratios are not reported. Instead the standard errors of these coefficients are reported*.

A complicating factor in the consideration of estimated models is the existence of correlations between the coefficient estimates. Clearly there exist strong correlations in both socio-economic and level-of-service variables in the basic data, and these tend to lead to correlations between the coefficients estimated for the correlated variables. For example, if travel time for two modes is positively correlated, then it is to be expected that separate coefficients for those travel time variables would also be positively correlated. These correlations in themselves cause no theoretical problems, as there is no requirement in the modelling for independence in the data, but they do cause the standard errors of the coefficient estimates to be increased. The logit estimation program was amended to print out the correlations in the estimates, but space precludes reporting the values. Few correlations with absolute values in excess of 0.3 were found.

---

* Special consideration also applies to "size" variables, and this is explained with the relevant models in Chapter 5.

The formal tests described above are derived from the inverse matrix of the second derivative of the likelihood function. Another series of tests can be derived from the values of the function itself. Because likelihood is a well-defined statistical concept, a series of standard tests and concepts can be used.

The most direct test available is the likelihood ratio test. If one model is a generalisation of another, then the likelihood is inevitably improved. Whether this improvement is significant can be tested by calculating the test statistic

$$2 \{ \log L_1 - \log L_0 \}$$

where L_0 is the basic model and L_1 the generalisation. Under the null hypothesis that L_0 is the correct model, the statistic is distributed χ^2 with the appropriate number of degrees of freedom. High values of the test statistic thus indicate significant improvements in fit obtained from the generalisation. For the generalisation of introducing a single variable this test is equivalent to the t-ratio test, but for more complicated generalisations this test may be the only one available .

The likelihood ratio test is formally available only when comparing models that are directly related in the manner described. Likelihoods may also be used, however, in a less formal manner to compare models that are less closely related but estimated on the same data. An appreciation of the magnitudes of likelihood changes that are significant in the formal likelihood ratio test is helpful in this less formal comparison of likelihoods.

A direct comparison of likelihood can be made only when the data on which the two models being compared are estimated is the same. It is sometimes necessary to compare models estimated on more or less different data, even in different studies, and for this purpose the ρ^2 statistic has been devised.

The statistic is explained by McFadden* . It is calculated by

$$\rho^2_{\theta_o} = 1 - \frac{L^*(\hat{\theta})}{L^*(\theta_o)}$$

---

* D. McFadden, Conditional Logit Analysis of Analitative Choice Probabilities in P.Zarembka (ed) Frontiers of Econometrics (1974).

where L^* is the log likelihood function evaluated at $\hat{\theta}$, the best estimate of the parameters, and at θ_o, some base value. It can be seen that ρ^2 will generally lie between zero and one, and that the better the model (i.e. the larger(less negative) that $L^*(\hat{\theta})$ becomes), the nearer will ρ^2 be to one.

For many models, it is suitable to take θ_o as being all zero values for the coefficients. The ρ_o^2 statistic thus defined will then give the improvement in explanation obtained by allowing non-zero values in the paramters. In other cases, when alternative-specific constants are present, it is more useful to take θ_o as being zero except for the alternative-specific constants. The ρ_c^2 defined in this way then gives the improvement relative to a model containing those constants.

This ρ^2 statistic is clearly strongly analogous to the R^2 statistic of linear regression. In fact, if normal distributions of the error terms are assumed, least-squares linear regression is exactly equivalent to maximum likelihood estimation of the parameters. In this case, a ρ^2 statistic can be defined (as above) for the regression, and this is exactly the same as the usual R^2 statistic.

Values are reported of ρ_o^2 and, where appropriate, ρ_c^2 for all the models presented in Chapter 5. These may be used to assess the overall quality of the models in comparison with each other and with models obtained in other studies.

CHAPTER 5. MODEL ESTIMATION

This Chapter presents the results of the estimation of the models. The complete system contains 37 models, all of which are listed in Table 6. Section 5.1 outlines the conventions of notation that are used, and the models themselves are then described in Sections 5.2 to 5.7, following the numbering scheme given in Table 6. The specifications of the models are given in Appendices.

5.1 CONVENTIONS OF NOTATION

The very large number of models estimated for this study requires us to adopt a consistent form of presentation. The models are presented in the following manner: the variables comprising a given model's specification are listed according to a mnemonic name, followed by a definition of the variable. The definition of each variable includes the following information:

- The manner in which the variable is calculated.
- The alternatives to which the variable applies. That is, whether the the variable is specific to a particular alternative, applies to a few (but not all) alternatives, or is generic across all alternatives. This information is equivalent to a list of the alternative-specific functions in which the variable enters.
- The level of applicability of the variable (whether the variable applies to a household, person, origin zone, destination zone, tour, zone type, etc.).
- The type of variable: i.e., if the variable is a size (or attraction) variable, a "logsum" variable, or a "constant". In the absence of an explicit indication of one of the above three types, the variable is understood to be a "typical" variable. Size variables will always be listed after all other variable types for each model in which they are included.

Following the definition of the variable, these tables give the estimated value of the coefficient of the variable and a measure of the accuracy with which this is estimated. In the "estimated coefficient" column, the maximum likelihood estimates of the coefficients are presented for "typical" and unconstrained "logsum" variables. The "constrained" value will be indicated for all constrained variables (logsum or size). If a "constant" has been adjusted (to correct for choice-based sampling, for example), only the adjusted value is shown and no measure of accuracy is given. For size variables, the estimated value is the natural logarithm of the coefficient multiplying the variable

Table 6. Component Models in the Model System

Model Type	Purpose	Report Section
Slow-mode Sub-mode Choice	WORK EDUCATION SHOPPING, PERSONAL BUSINESS SOCIAL RECREATION MISCELLANEOUS	5.2
Car-mode Sub-mode Choice	WORK SHOPPING, PERSONAL BUSINESS SOCIAL, RECREATION MISCELLANEOUS	5.3
Main Mode Choice	WORK EDUCATION	5.4
Joint Mode - Destination Choice	SHOPPING PERSONAL BUSINESS SOCIAL RECREATION MISCELLANEOUS	5.5
Destination Choice	WORK, USUAL WORKPLACE WORK, UNUSUAL WORKPLACE EDUCATION	5.6
Frequency Choice (STOP/REPEAT)	WORK, USUAL WORKPLACE WORK, UNUSUAL WORKPLACE SHOPPING PERSONAL BUSINESS SOCIAL RECREATION MISCELLANEOUS	5.7
(1,2+)	EDUCATION, STUDENTS	5.7
(0,1+)	WORK, USUAL WORKPLACE WORK, UNUSUAL WORKPLACE EDUCATION, STUDENTS EDUCATION, NON-STUDENTS SHOPPING PERSONAL BUSINESS SOCIAL RECREATION MISCELLANEOUS	5.7

NOTE: the numbering of the models in the Appendices follows the numbering of the report sections.

in the compound size measure (i.e. constrained size variables have coefficient log rithms of 0.0).

The measure of accuracy of estimation normally reported is the absolute value of the "t_0-ratio", indicating the significance of the difference between the estimated value of the coefficient and zero. For logsum variable coefficients, both the t_0-ratio and the absolute value of the "t_1-ratio" are shown (the t_1-ratio indicates the significance of the difference between the estimated value of the coefficient and one). In the case of unconstrained size variable coefficients, t-ratios are not shown;in their place is the standard error of the estimated value. Of course, no "accuracy of estimation" information exists for the coefficients of the constrained variables.

After the variable-specific information summary statistics for the model as a whole are given. These statistics incluce:

- for each alternative, the number of observations with the alternative available, and the number of observations observed to have chosen the alternatives,
- the total number of observations,
- the values of $L^*(0)$, $L^*(C)$, and $L^*(\beta)$, and
- the values of ρ^2, ρ_c^2,

In some cases, not all of these values are presented. These statistics are explained in Section 4.6.

5.2 SLOW-MODE SUB-MODE CHOICE

This section of the report will describe the models of choice among the three "slow" modes of walking, bicycling, or using a moped, given that the traveller makes his choice among the slow modes. Together, these slow modes accounted for almost two-third of all observed tours in the Zuidvleugel household interview data set. Six models were estimated, one for each of six purpose groups.

These slow-mode models are not intended to be sensitive to the wide range of policies that might be used to influence slow-mode choice (e.g., the construction of separate right-of-way for bicycles); data required for such an analysis would have included more detailed information of bicycle paths and signalization, and more accurate information on precise locations of origins and destinations of tours than is available for the Zuidvleugel study. The pur - pose of the slow-modes models is to provide the basis for constructing a va-

riable (sensitive to a variety of socio-economic, distance and locational descriptors) measuring the attractiveness of the slow modes as a group, for use in models of (main) mode choice. This information, incorporated in the logsum variable, allows an improvement in the explanatory power of main mode (or joint) choice models. The models are also of interest in themselves in indicating the characteristics of the users of each of the slow modes.

5.2.1 TOPICS IN SLOW-MODE CHOICE MODELLING

In this section, a variety of topics common to all six slow-mode choice models are discussed. These include the data sets used to estimate the parameters of the model, a series of experiments to find the most appropriate measure of distance among the several possibilities, a series of tests to find the most appropriate set of purpose groupings, and a discussion of the treatment of moped availability. The reasoning behind the omission of zonal models of slow-mode choice is also given.

Estimation Data Sets

The data used for slow-mode modelling consists of the tour files derived from the household survey, distance measures between households and destinations derived from network analysis, and a small amount of zonal information. Because the size of many of the zones in the Zuidvleugel study area is large compared to the trip lengths of most slow-mode trips, the zonal data is expected to be of less value than in models of other travel choices. The most accurate measure in the network files provided a slow-mode distance from the location of the household to the centroid of the destination zone. Because of the relatively large zone sizes (and because household locations are coded only to 500 meter accuracy), this measure of distance is less precise than desirable for slow-mode modelling. Distance was used as the sole level-of-service measure on the assumption that congestion is not an important determinant of slow-modes speed*. Of the 11,002 valid home-based tours potentially available as observations, 6,801 are slow-mode tours. After a variety of deletions (details are given in Table 7), a total of 6,753 tours were used in estimation.

---

* While some information about the variation in bicycle speeds is available for different links in the network, it was judged that the marginal value of this information would not balance the inconvenience of processing the more complex network.

Table 7. Slow-Mode Sub-Mode Choice Data Sets

Purpose	Purpose Codes*	Number of Valid Home-Based Tours	Deletions					Observations Modelled
			No Valid Chosen Alternative	Miscoded Household Location	Trip Length Measurement Error	Other Deletions**	Total Deletions	
Work	24,25	733	0	0	0	0	0	733
Education	30	2259	4	5	29	1	39	2220
Shopping & Personal Business	26,28,29	1551	0	2	0	0	23	1549
Social	27	625	1	0	0	0	1	624
Recreation	31	708	1	1	0	0	2	706
Other	32-36	925	0	4	0	0	4	921
Total	-	6801	6	12	29	1	48	6753

* For purpose code definitions, see Zuidvleugel Study Report 5.

** Other deletions typically consist of outlying observations with undue influence on parameter estimates.

Measures of Distance

A series of tests were performed to compare four alternative measures of distance: (1) network distance between the household location and the centroid of the zone containing the tour's destination; (2) straight-line distance between the locations of the household and the tour's destination, both measured to 500 m accuracy in the coordinate system; (3) straight-line distance between the household location and the centroid and of the zone containing the tour's destination; and (4) straight-line distance between the centroids of the zones containing the tour's household and destination. Network distance between the zone centroids was available but was not tested. The test procedure compared the values of the log-likelihood function (at convergence) for pairs of models estimated with distance measures in these four different ways. Although the second measure proved best in terms of statistical significance, it was discarded since it was not possible to use the actual location for the destination choice models (which had to be on a zonal basis), and that is was desirable to have comparable accuracy for all the different model types. Therefore, the first measure was selected because it was possible to use it in all the models and the results which showed it to perform second best of the four distance measures.

Slow-Mode Purposes

It is naturally unreasonable to estimate different models for each of the 13 different purposes recorded in the survey. Apart from the complexity that this would cause for the estimated models, there is insufficient data for some purposes to make separate estimation feasible. Some degree of merging of purposes is therefore required. In some cases (e.g. when the number of tours was very small) this merging was done arbitrarily, but in other cases, a goodness-of-fit test was performed to check the acceptability of modelling different purposes with the same model.

The objective of this test is not the strict statistical question of whether the coefficients for the model for one purpose are exactly the same as those for the other. First, this is because we have no reason to suppose that they are the same, and we would expect that a large enough data set would reveal a significant difference. Second, however, we do not have a very large data set, and the large number of parameters in the models means almost certainly that we will not detect the difference that we believe exists. The test is rather the less formal one: is the difference in the coefficients worth the trouble of estimating and applying separate models?

Tests were made for the combination of the two work purposes and for combinations of any pair from social, recreation and "other" tours. The test compared the predictions given by merged models with those given by models estimated on the separate purposes separately. The results of these tests are shown in Table 8. For the work merging, too few unusual workplace tours used slow modes to permit development of an independent model, and the test was simply that the inclusion of these tours did not drastically affect the model for usual workplace tours. Table 8 shows the log-likelihood loss for usual workplace tours was only 2.2, and the merged model was accepted. For social and recreation, however, the two purposes lost 18.7 and 19.6 respectively on merging, so this merging was rejected. Further, the loss on merging these purposes with "other" tours was 15.5 and 14.6 respectively, to which must be added the loss (not calculated) of the "other" tours themselves, so that these mergings were also rejected.

These tests are, of course, somewhat informal, since they require a feeling for what is a "large" likelihood loss. But since, for the reasons explained, no formal tests can be done, the informal tests appear appropriate. Fortunately, the results were rather clear as to which purposes should and which should not be merged.

Table 8. Results of Purpose Merging Tests, Slow Sub-Mode Choice

	Purpose		
	Work Usual Workplace	Social	Recreation
Log-likelihood value in separate model	-417.6	-430.4	-441.7
Log-likelihood of observations in merged model work, usual workplace; with work, unusual workplace	-419.8		
Social; with recreation		-449.1	-461.3
Social; with other		-445.9	
Recreation; with other			-456.3

Moped Availability

The availability of the moped mode is an interesting issue. For the purpose of the models, it was assumed that all individuals who could legally use a moped (age 16 or over) had that mode available. The most likely alternative assumption would have been to condition moped availability on moped ownership by the household. However, such a course would have led to the requirement that either: (1) future moped ownership would have to be exogeneously predicted; or (2) an explicit model of moped ownership would have to be developed. Regardless of which course of action might be selected, it was judged that the additional complexity introduced would not be justified by the potential for improved explanatory power or behavioral representation.

Several variables appear in the models that are best interpreted not as components of utility but as conditioners of availability. As discussed in Section 4.2 these variables appear in a strictly improper form, but the errors thus introduced are small since only a small minority actually choose moped. In particular, the assumption was made that anyone whose best mode for work or school was moped would simply buy a moped, but that the use of a moped for other purposes was conditioned by use both by the household by the individual for work or school purposes.

Omission of Zonal Models

The slow-mode choice models were developed in the fully disaggregate form. The variables that appear in them are primarily the type of socio-economic variables that are not available in the aggregate models. Further, the level-of-service measure (distance) become much more inaccurate for these very short tours when we move from fully disaggregate to zonal models. Therefore, it is reasonable to expect that the quality of zonal slow-mode choice models would be greatly reduced relative to the fully disaggregate ones.

The slow-mode choice models, however, are of somewhat less interest for zonal forecasting, where the interest is more in regional, long-term issues than in the type of problems for which slow-mode models are helpful. For these reasons, slow-mode models were not developed at a zonal level but only at the most disaggregate level.

5.2.2 The Work Tours Model

The work tours model predicts choice among the slow modes for work tours to both the usual and unusual workplaces of individual travellers. As in all the slow sub-mode models, bicycle utility has been arbitrarily set to zero; thus,

the coefficients of the variables in the walk and moped utility functions must be interpreted relative to their implicit value of zero for the bicycle mode*. The hypothesized determinants of slow-mode choice were the impacts of life-style differences, systematic taste variation across different types of individuals, other types of preferences that may be reflected by socio-economic variables and the effects of distance.

In Appendix A, model A2.2 is the model for slow-mode choice of work tours. Looking first at the distance variables (number 2, 3, 13, 14), it is clear that as distance increases, moped is more preferred than bicycle, and walk is less preferred than bicycle. This result is expected, of course, as it is consistent with one's prior expectations about the relative speeds of the three slow modes. One also notes that the marginal differences in utility decline as distance increases, partially(possibly) because of inaccurate distance measures**.

Figures 3 and 4 display how the distance variables affect utility, all other things remaining equal. Figure 3 is constructed by using the variables appearing in the walk-mode utility function (variables 2 and 3 in the model). Note that for one-way tour distances between zero and one kilometer, only variable 2 and its coefficient are included in the calculation of utility; for distance over one kilometer, the utility is computed with both variables and their respective coefficients. The selection of the value of one kilometer as the appropriate "break point" was determined empirically on the basis of goodness-of-fit, within the bounds of theoretically reasonable values. Figure 3 presents, for the moped mode, the same information that Figure 4 contains for the walk mode. As expected, greater distance travel translates into increasing preference for moped use over bicycle use; the rate of increased preference decreases at distances over nine kil meters. The break point of nine kilometers was selected in the same manner as the one kilometer break point for the walk/bicycle tradeoff. Note that the slope of the graph is steeper in Figure 3 than in Figure 4, so that for longer trips walk is almost never selected.

---

* There being no variables specific to any one of the modes, we can only assess parameters relating to the differences between them. Thus we might as well set the parameter for one of the modes zero for each variable. This has been done for every variable for the bicyle mode for simplicity, because the mode occupies a central position relative to walk and moped and (for technical reasons) because it is the most frequently chosen.

** Because slow modes are short trips, the assumption that all tours travel to zone centroids may exaggerate distances and obscure much of the distance variation between modes, particularly in the larger zones.

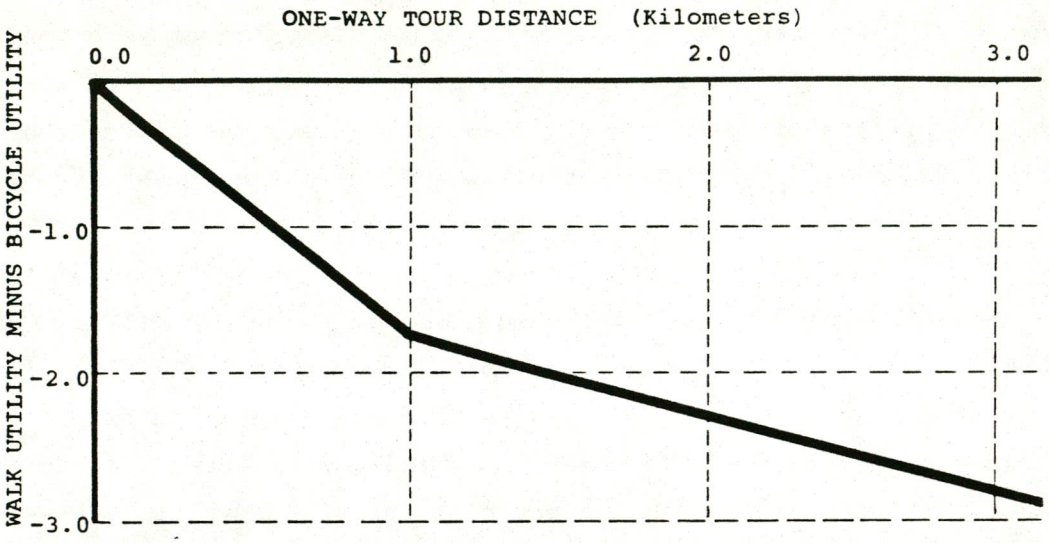

Figure 3. Walk Minus Bicycle Utility vs. Distance, Disaggregate Slow Sub-Mode Choice Model for Work Tours

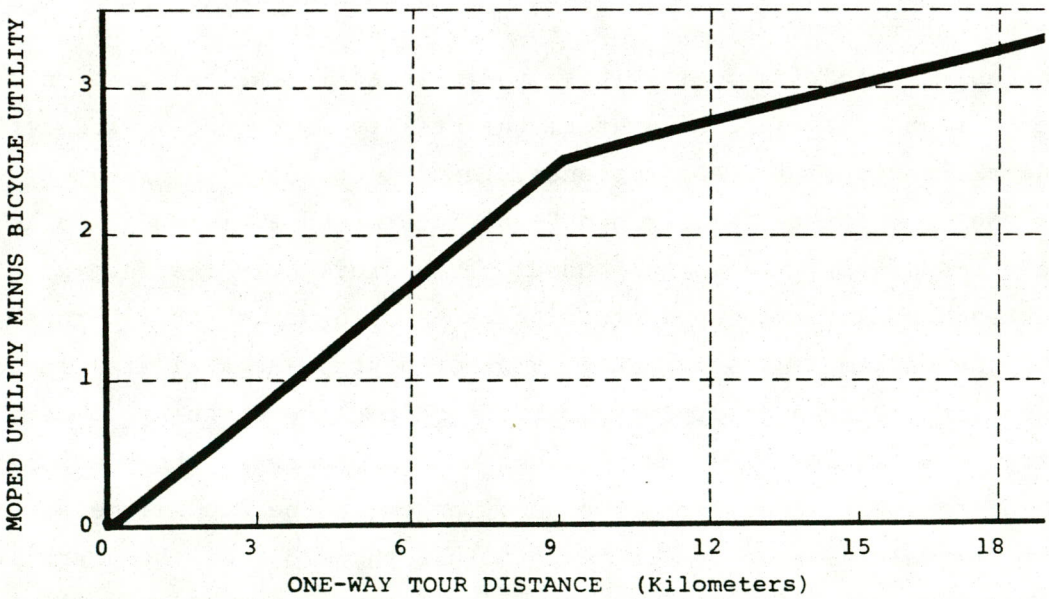

Figure 4. Moped Minus Bicycle Utility Vs. Distance, Disaggregate Slow Sub-Mode Choice Model for Work Tours

The other variables in the model specify a mode preference by certain socio-economic groups. These variables show that there are significant differences in slow-mode choice based on age, occupation, and time of day. These variables also capture moped availability and ownership effects. Age seems to be important mainly for the younger workers, who prefer mopeds and dislike walking relatively more than the general population. The moped variables here may partially be a proxy for a car availability effect as younger people are less likely to have access to a car and may use mopeds as substitutes for cars to some extent. There is also an effect for the middle age group (25-45 years), showing a preference for bicyle over moped*. One hypothesis for this effect is that people in the range of 25-45 years would use a car if they used a motorized mode, and bicycle has certain advantages (e.g. exercise) over moped, if a slow mode is used. Older people may be less inclined to drive and generally less able or willing to use bicycles, and may come from a generation that find the use of mopeds an acceptable compromise.

Time of day effects are captured in two different variables in this model: W-LUNCH and M-EARLY. These variables represent entirely different things. The W-LUNCH variable suggests that people who return home for lunch from work are less likely to walk, probably because lunch time is limited and the additional speed of a bicycle is an advantage. It could be expected that such an effect would be found, and the variable is therefore retained although it is not significant. The M-EARLY variable, which is much more important, is probably measuring a variety of effects. The making of early work trips is correlated with profession so it may be a profession (or status) effect which is not fully captured by the profession variables. Alternatively, it could be an effect of speed being more important early in the morning. Another potential cause is a darkness effect, since it is not light until after 8:00 during the surveyed months.

Occupation variables enter the utility functions in several places. These variables also represent many different effects. The moped variables are probably largely measuring a "status" effect; mopeds may be viewed as an inferior mode for social reasons by some groups. Walk effects could be functions of inaccurate distance measures in the case of agricultural workers (due to larger zone sizes); for other occupations (service, commercial), the walk variable may reflect decreased bicycle access due to congestion in city centers.

---

* However, as age increases further, moped preference increases again.

The W-INC-HI variable has a very significant negative impact on walk utility (although a higher value of time could explain part of this, it is unlikely to have so large an impact). The lack of a driving licence has a large impact on moped usage, due to the increased attractiveness of moped when the car driver mode is unavailable. The NON-FIXED variables are included, even though their significance levels are low, to insure against biases introduced by including tours with non-fixed workplace in the model: these variables should be seen more as corrections to the constants for non-fixed work tours rather than as new variables in their own light.

5.2.3 THE EDUCATION TOURS MODEL

This model predicts slow-mode use for tours made for education purposes. As was the case with the work models, it was presumed that distance and socio-economic characteristics would capture the most important determinants of choice; in addition, the educational status of travellers was also hypothesized to play a major role in the decision.

Table A2.3 in the Appendix gives the recommended model for slow-mode choice for education tours. Age and education status are clearly the most influential effects for the choice between walk and bicyle. Age is not so significant for the choice between moped and bicycle, primarily because of the low degree of variability (moped use is restricted to those who are 16 years of over). Because education trips are disproportionately short, even by slow-mode standards, the distance measures are particularly difficult to obtain, as is shown by the complicated forms of the variables (2,13, and 16).

The education status variables present some difficulties of interpretation. For the walk/cycle split, five groups were found: primary, two classifications of secondary education, higher education and those who had no data for education. It appeared that those who gave no answer behaved most like the higher secondary education class, having a relative preference for walk, compared with the primary and higher education groups. The lower secondary group had an even greater preference for walk over cycle. For the moped/cycle choice, those with advanced secondary or "other" education (the latter largely apprenticeships, etc.) had a strong relative preference for moped.

The income-not-answered variable was retained to avoid biases. Because of the high frequency of non-reponse to the income question, we were unwilling to discard observations missing this data.

5.2.4 THE SHOPPING/PERSONAL BUSINESS TOURS MODEL

Like the work and education models, the working hypothesis for analyzing shopping and personal business tours was that slow-mode choice was primarily determined by distance and systematic personal taste variations (as measured by socio-economic characteristics). Additionally, it was supposed that moped availability/ownership measures (in particular, the use of mopeds for work or education tours) would be important, and that density measures would refect the opportunities for making the short trips that typify slow-mode use.

Table A2.4 in the Appendix presents the recommended model of slow-mode choice for shopping and personal business tours. Most of the variables in the model are similar to ones descibed in the previous models, but in addition, this model contains measures of moped availability: M-W, E-M and M-W, E-M, HH (variables 27 and 28 in Table A2.4). These variables are included on the premise that a joint moped ownership/usage model is not particularly appropriate for the non-work and non-education purposes. Thus, although it was assumed that a moped could realistically be expected to be purchased for work or education travel, it was decided that this is a less appropriate assumption for other purposes. The additional moped availability measures are intended to capture the increased probability of using a moped given that one is owned by the household (for another reason). The variables conditioning moped use for this purpose on moped use for work or education are therefore measures of moped ownership, but avoid the necessity of developing specific models of moped ownership. The results of the estimation indicate that there are, as expected, significant effects that can be captured in this way.

Another set of variables which figure more prominently in this model than in previously reported ones are the various density measures (variables 3, 4 and 30). It was expected that the use of the "slower" slow modes (especially walk) would increase in dense zones. Travel in dense zones was also observed to lead to an increased preference for moped. A possible explanation for this observed effect is that a closer link exists between choice of car or moped than between car and the other slow modes. Thus congestion effects may reduce car usage leading to a relative increase in moped use. Such a hypothesized model structure was not, however, tested.

The distance variables (2 and 18) appear in this model in a simple form. Age variables (5, 7, 8, 22, and 23) and the "retired" variables (6 and 26) are

important in this model. These variables indicate a relative preference for walk among young adults and old people, offset by the "retired" variable, and reluctance to walk by children; mopeds are relatively preferred by the 16-20 age group, and to a lesser extent by the 30-45 group and retired people. Income variables are less important. Time variables (10, 11, and 14) show an interesting pattern, bicycle being relatively preferred in the early evening, but not in the late evening. The NHB variables (13 and 29) show a preference for faster modes if a detour must be made. Corrections to the constants for personal business tours have been retained. None of the variables appear to have anything other than a natural value that might be expected.

5.2.5 THE SOCIAL TOURS MODEL AND THE RECREATION TOURS MODEL

The choices among the slow sub-model for social tours and for recreation tours were hypothesized to be sensitive to the same types of variables used in the shopping and personal business model: distance, density variables, time-of-day, socio-economic characteristics, implied slow-mode preferences based on observed choice of slow mode for work and education tours, and variables indicating the occurrence of detours.

Table A2.5 in the Appendix shows the results of the selected models of social and recreation tours*. The results shown for the social and recreation models are not as good as those for the previous models. Distance coefficients, in particular, are not strong, and for distances under one kilometer, the signs are unreliable.

Few new variables are contained in these models, although the ranges of certain categorical variables have been changed as appropriate for these tour purposes. For instance, the definitions of time-of-day variables have been changed because of the increased number of late night tours. The walk-to-work variable was included to capture any taste variation (preference) that may exist among certain individuals for walking relative to bicycling. It was felt that if walk was used for work, this would be a measure of preference for walking for other types of equivalent distances. This proves to be the case in the social model, but the coefficient is not significant though appropriately signed in the recreation model. Otherwise, the results are satisfactory in general.

---

* Note that the specification of the social and recreation models were identical (in the interest of simplicity of later applications). This led to the inclusion of a number of statistically insignificant variables in both models. The models are both presented in Table A2.5.

5.2.6 THE SLOW SUB-MODE CHOICE MODEL OF MISCELLANEOUS TOURS

The model of miscellaneous tours is a grouping together of all the purposes that were not included in the previous models. There is no objective reason why these purposes should be grouped together into one model. However, it was not appropriate to group them with other purposes and a set of six sub-mode choice models was judged to be as cumbersome as is acceptable. Although the results of this model are not particularly important in their own right (because of the diversity of the tour purposes included), tours for these purposes account for too large a percentage of the total slow mode tours to be ignored so some means of allocating these tours to slow modes was needed.

The model is presented on Table A2.6. The specification for the model is essentially similar to those of the social and recreation models. Again, this was done in the interest of simplicity, but has the consequence of leaving some extraneous and/or (statistically) insignificant variables in the model. For example, the time-of-day variables for the walk-cycle choice of modes for these purposes to those used for work and education. In summary, however, the model seems to fit rather well considering the variety of purposes involved.

5.3 CAR SUB-MODE CHOICE

In this section, models of the choice between driving a car and travelling as a car passenger are presented. The models represent choices for all purposes except education tours - in the case of education tours, the choice between car driver and car passenger was analyzed in the (main) mode choice models, described in Section 5.4. The remaining purposes have been grouped into four purpose categories (work, shopping/personal business, social/recreation, and serve passenger/miscellaneous); separate models have been developed for each of these four purpose categories.

There are two main functions served by the car sub-mode choice models: they provide a method for predicting average car occupancy; and they yield an important measure of the attractiveness of the car mode (the logsum variable) for use in the (main) mode choice models.

5.3.1 TOPICS IN CAR SUB-MODE CHOICE MODELLING

A number of issues were encountered during the process of developing the car

sub-mode choice models. First, a series of structural issues were considered, together with the definition of the proper set of alternatives. These structural and "choice set" topics were closely related to the question of the accuracy of data available to analyze the choice of car sub-mode. The representation of the availability of alternatives played an important part in the decisions made on how to represent car sub-mode choices for education tours. Finally, as with the slow sub-modes, the issues of dividing the data into purpose groups had also to be solved for the car sub-mode choice models.

Model Structure and Data Issues

In the household survey responses four categories of car usage were distinguished: (1) car driver alone, (2) car driver with passenger, (3) car passenger in a car owned by the household, and (4) car passenger in a car not owned by the household. To explain the full richness of the data, a car sub-mode choice model should similarly distinguish these four fundamental alternatives.

Because of the potential difficulty of testing alternative model structures empirically, it is useful to use other criteria for the structure of car sub-mode choice. In particular, the quality of the household survey data influences the model structure that can be selected. This is true because there is evidence that the home interview survey data is unreliable in the distinction between driving with or without passengers, and the distinction between passengers in a household car or nonhousehold car may similarly be unreliable. The errors in the driving distinction are suggested by a contradiction between two methods of calculating car occupancy. One method is to calculate the ratio of the number of all car travellers (drivers and car passengers) to the number of car drivers, giving 1.38 persons per car. An alternative is derived from the ratio of twice the number of drivers with passengers plus the number of drivers without passengers divided by the total number of drivers: the result is 1.48 persons per car. The contradiction can have two causes: (1) errors by some respondents in reporting that they drove with passengers when in fact they drove alone, and (2) similarly miscoded data in the distinction between drivers and passengers. However, the design of the questionnaire is such that respondents' errors of driver type are more likely than respondents' errors in the distinction between driver and passenger. Because the questionnaire design is similar, a similar problem of respondent error might have arisen for passengers in the distinction between cars owned by their household and other cars, but there is little direct evidence

of this. Cross tabulations of the data set reveal that some individuals who report that they were passengers in a household car belong to households which report owning no cars, but the exact severity of the problems with the passenger mode information is subject to uncertainty.

Because overcoming these data problems directly would have been very difficult, it was decided not to model the four-alternative car sub-mode choice process, and to concentrate our attention on the binary choice between car driver and car passenger. Potentially lost is a richer understanding of the process that car users go through in their choice of car sub-mode, although the ability to forecast car occupancy is retained.

Data Sets for Analysis

The data used to estimate the car sub-mode choice models have two sources: the tour data file and the set of zonal data files. Table 9 shows how the final set of observations were derived from the original household survey data. Clearly, the majority of deleted observations are tours made by car passengers where the traveller did not have the driver alternative available because either the household owned no car, or (usually) the traveller did not possess a licence to drive. Since these travellers had no choice between car sub-modes, they give essentially no information for this model and are therefore omitted from the data on which the model is estimated. Further investigation revealed that most of the car passengers who are not licensed to drive are children who are too young to drive. As can be expected, the problem of these deletions of children is most important for the education tours, where 96 percent of the car passengers could not be used for model estimation.

Availability of Alternatives

In the early stages of developing model specifications, tours were eliminated from consideration if individuals reported that a car was not available in the time period when the tour was begun*. However, when this "availability constraint" was tentatively removed, it was discovered that among the newly-introduced tours were many driver tours. Thus, there are apparently a number of people who have reported both that a car was not available for their use for a particular tour and that they drove a car for that tour. This anomaly indicates a structural problem with availability data collected (as in the Zuidvleugel survey) for specific time periods: that the times at which a car is available are most unlikely to match exactly the periods specified on the

---

* Again, on the hypothesis that these people had no choice.

Purpose	Purpose Codes	Number of Valid Home-Based Tours*			Observations Deleted							Observations Modelled			
		Driver	Passenger	Total Car Users	No Car or No License	Invalid Zone Origin or Destination	Miscoded Occupation	Impure Mode	outlier	other	Total	Driver	Passenger	Total	%
Work	24,25	1101	165	1266	132	25	-	4	1	1	161	1034	71	1105	87
Shopping, Personal Business	26,28, 29,35	380	158	538	130	5	4	2	-	9	150	351	37	388	72
Social, Recreational	27,31	400	317	717	261	25	6	3	-	40	335	337	45	382	53
Service Passenger, Miscellaneous	32-34, 36	388	104	492	60	6	4	2	-	47	119	348	25	373	76
Subtotal	-	2269	744	3013	583	61	14	11	1	95	765	2070	178	2248	75
Education	30	63	139	202	109	-	-	-	-	24	133	63	6	69	34
TOTAL	-	2332	883	3215	692	61	14	11	1	119	898	2133	184	2317	72

*From a total 11,002 tours for all modes.

Table 9. Car Sub-Mode Choice Data Sets

survey form, and that ambiguities will inevitably arise. For this and other reasons discussed in Section 4.2, it was decided not to use this availability data. Instead, car availability was represented simply by household car ownership and the driving licence holding of individuals.

However, even the simple requirements of car ownership and driving licence status can eliminate many travellers from a car sub-mode choice, particularly for education tours. In the case of education tours the effects were severe enough so as to preclude the ability to develop a model. For this purpose, the data set is not large, even initially, and the deletion of the 66 percent who do not have the choice of driving reduces it to a dangerous level for reliable modelling. Further, Table 9 shows that slightly over 31 percent of the tours made by car for education purposes were made by driving. However, if only the sub-set of education travellers who use the car mode and could drive are considered, then over 91 percent drive. In the car sub-mode choice model, only the latter group can be considered - a group composed exlusively of individuals licenced to drive whose households own at least one car.

For the above reason, an education sub-mode choice model is not recommended and the driver/passenger choice was analyzed in the (main) mode choice model. A model was, however, estimated, and this model is presented in the following sub-section.

The Education Model

Table A3.1 in the Appendix illustrates the most satisfactory model that was found for this purpose. Perhaps its most immediately striking feature is that only three significant variables were found. Note also that aside from parking costs, the model incorporates no measures of the transportation system or its level of service. Furthermore, the importance of the travellers' "other" educational status is behaviorally unclear at best. Nevertheless, the model's overall explanatory power must be viewed as satisfactory. Parking cost, the traveller's sex and education status together explain about 20 percent of the uncertainty remaining in the choice between being a car driver and car passenger, after accounting for the alternative specific constant.

The model is strictly applicable only to individuals similar to the 69 travellers who comprise the estimation data sample: adults licensed to drive, living in car-owning households, and making education tours by car. Unfortunately, as has been discussed, the large majority of travellers making

education tours by car (or any other mode) are children who are not licensed to drive. Application of the model to children and other individuals systematically excluded from the sample is extremely unreliable. For this reason, too, the model was abandoned in favour of a (main) mode choice representation of the choice among car driver, car passenger, public transport, and the "slow" modes (walk, bicycle, moped), in which the education travel decisions made by children can be incorporated along with those made by adults. This model is described in Section 5.4.3.

Car Sub-Mode Choice Purposes

As implied by Table 9 , a considerable amount of merging of tour purposes was found to be acceptable when modelling car sub-mode choice. Tests similar to those described in Section 5.2.1 were applied for these tours also to test the appropriateness of developing combined models for: (1) work at usual workplace and work at unusual workplace; and (2) shopping and personal business. Tours with social and recreation purposes, and tours with serve-passenger and miscellaneous purposes, were merged on the basis of sample size and other prior considerations.

5.3.2 THE WORK TOURS MODELS

Table shows that most individuals making a work tour by car choose to drive rather than ride as a passenger. Thus, although almost 70 percent of the initial log likelihood is explained by the model, the first 66 percent of it is explained by the aggregate shares. The model reflects clearly a number of effects that influence car sub-mode choice strongly. For example, as household car ownership increases, a traveller's propensity to drive rather than ride as a car passenger also increases. Male travellers are more likely than females to drive, given that the car mode has been selected. As the cost of car travel increases (as measured by greater distances and higher parking costs), a car traveller's probability of being a passenger, and thereby sharing the costs, increases*.

The "parking difficulty" variable (D-PR-DFC9) was originally included in the model's specification to reflect the hypothesis that if a traveller makes a tour to a destination in which parking is difficult (at a time when parking

---

* Although a car driver might have passengers and therefore be able to allocate the travel costs over more than one person, a car passenger will always be travelling with at least one other person, and perhaps with more than one other person.

is difficult), he or she will be less likely to drive so as to be more likely to be able to "share" the cost of difficult parking with other individuals. However, the empirical result of the positive sign of the coefficient of this variable that such a traveller is more likely to drive. The explanation for this result may be due to the fact than the "parking difficulty" variable can also be thought of as indicating a city centre destination together with a late arrival time. It may be that the workers who arrive at their destination by car after 9:00 am in central areas are high income "white collar" managers or executives, who drive instead of ride as a passenger for reasons which may not be related to level-of-service. Alternatively, these workers may have free parking places permanently allocated to their exclusive use. Several attempts to find socio-economic variables to describe this effect more directly were not successful.

The density variables (P-HH-DEN and P-EMP-DEN) were originally included in the specification in order to reflect the hypothesis that as the densities increased, the ease of sharing rides would be increased, thereby resulting in a greater likelihood of riding as a car passenger compared to driving. The empirical results show, however, that as the densities increase the probability of car driving increases. These results may partly be caused by a correlation of other-mode usage with household and employment density (as densities increase, (1) the quality of public transport service increases, leading to public transport usage, and (2) trip lengths decrease, leading to slow-mode usage). They may also be due to unrepresented socio-economic variables.

Zonal Model

The specification of the zonal model is a slightly altered form of the disaggregate model. Referring to Table B3.2, an alternative measure of car ownership (zonal car registrations divided by zonal population) was used, resulting in a reasonably accurately estimated coefficient. Because activity time at the primary destination is not available for aggregate estimation (or forecasting), the parking charge used is the hourly rate rather than the actual cost. A new variable, P-MORN-EP, was also introduced to indicate a propensity to be a passenger if travel is in the peak periods of the day. Other variables (and coefficient estimates) are substantially similar to the fully disaggregate model presented in Table A3.2.

5.3.3 THE SHOPPING/PERSONAL BUSINESS TOURS MODELS

The estimated model is shown on the Table A3.3; of the seven variables, only

R-TT (passenger's travel time) and P-PB-CONST (a personal business tour constant) are new. As is often the case, travel time and travel distance are too closely correlated to permit the reliable estimation of their separate impact on the car sub-mode choice. The D-OTH-EDUC variable is an artefact which proved to have an important effect in the context of a model of shopping tours only; it was retained in the "pooled" model of shopping and personal business tours despite its low accuracy of estimation merely because its inclusion does not adversely affect the resulting model.

Notable in the estimation are that, unlike the model of work tours, (1) the impact of household density is as originally postulated;(As household density increases so do one's opportunities for ridesharing, resulting in an increased probability of riding as a car passenger.) and (2) the impact of parking difficulty is as originally postulated.(Travel to a zone with a constrained parking situation is accompanied by an increased probability of travelling as a car passenger, so that the "difficulty" can more likely be shared among multiple car occupants). Note also that the low level of statistical significance of the coefficient of P-PB-CONST variable can be interpreted as indicating that there is no evidence to suggest that personal business tours are fundamentally different from shopping tours (in ways that are not incorporated in other variables already included in the model.

Zonal Model

The zonal car sub-mode split model for shopping and personal business tours, presented in Table B3.3, is almost identical to its disaggregate version. The only difference apart from the use of zonal level data for travel time is that the D-OTH-EDUC variable has been omitted. Given the similarities in both data and specification, it is not surprising that the estimated coefficient, their t-ratios, and the models' summary statistics are all very nearly identical*.

5.3.4 THE SOCIAL/RECREATION TOURS MODELS

In the social/recreation car sub-mode choice model, presented in Table A3.4, only a few new variables can be found. For example, D-PC-RECA, the proportion of the destination zone area devoted to recreational use, reflects the hypothesis that travel to zones which are heavily oriented to recreation (wooded,

* The indentical results confirms the lack of consequence of including or excluding the D-OTH-EDUC variable in the disaggregate version's specification.

lakes, etc.), will not present the congestion and parking competition difficulties associated with travel to urban zones and so incentives for ride-sharing are less than otherwise. As the number of other vehicles owned by the household (D-#-OVEHS) increases, the competition for the family cars decreases, and so the probability of driving is found to increase. A similar effect is found from an examination of the "cars per family size" variable, D-CARS/HHS.

Zonal Model

The zonal social/recreation model specification (see Table B3.4)is a close approximation of the disaggregate model. In place of cars per household member, the variable cars per person (calculated on a zonal basis) has been substituted. The number of other vehicles had to be dropped, as had the education-related variables. The hourly parking rate was used in place of the parking charge. When compared to the disaggregate specification (Table A3.4), the coefficients of the measures of car ownership and parking cost are less accurate, while most other coefficient estimates are unchanged or improved. Overall explanatory power is slightly lower, as can be expected with the less precise aggregate data used for estimation.

5.3.5 THE SERVE PASSENGER/MISCELLANEOUS TOURS MODELS

The disaggregate car sub-mode choice model for tours which serve passengers or are made for miscellaneous purposes (medical travel, trips to church, and "other" purposes) is shown in Table A3.5. As with the other car sub-mode choice models, the results indicate that increasing household ownership of vehicles (both results indicate that increasing household ownership of vehicles (both cars and other vehicles))tends to increase driving as opposed to riding as a car passenger. Male travellers are also more likely to drive, whereas travellers with "other" educational status and travellers making long trip are more likely to be a car passenger. Travellers making any part of their tour during peak periods show a counter-intuitive increased propensity to drive, this applying primarily to medical, church, and serve passenger travel (when a similar model was estimated with data excluding the "other" purpose tours, the coefficient of P-PK-TRIP was found to be -3.08 instead of -0.561 as shown in Table A3.5.)The negative sign associated with P-PK-TRIP'S coefficient must be considered counter-intuitive due to the hypothesis that travel during the peak periods is more onerous than travel at other times during the day, and that therefore there is an incentive to "share" the additional disutility of peak period travel. A possible explanation for the negative

sign is that a majority of car travel for serve passenger, church, medical, and "other" travel must (for institutioanl reasons) occur at least partially during a peak period of the day. Since the majority of such travel is also accounted for by drivers, such an effect would necessarily result in a negative sign for the P-PK-TRIP variable.

Zonal Model

The zonal model was specified with a car driver constant term and a utility term constructed to correspond are closely as possible to the specification and results of the fully disaggregate model: the variables D-#-OVEHS and P-OTH-EDUC are omitted because their measures are not available at the zonal level of analysis; a D-MALE variable is multiplied by its disaggregate coefficient (see Table 22), 1.63; a cars-to-population ratio is substituted for the disaggregate cars per household size measure and multiplied by the cars per household size coefficient, 1.93; zonal travel time is substituted for disaggregate origin-destination travel time and multiplied by the origin-destination travel time coefficients, 0.0178; and a peak trip variable for passengers is multiplied by its disaggregate coefficient, -0.561. Theoretical considerations require the estimated coefficient of the constructed utility term to lie within the 0.0-1.0 range, with a value of 1.0 indicating no loss of accuracy in the aggregate formulation when compared to the disaggregate model. Table B3.5 shows the results of the aggregate model. The coefficient of the utility term is 0.706, well within the required range. Nevertheless, a comparison of the summary statistics of the disaggregate and zonal models (Tables A3.5 and B3.5) shows a substantial loss of explanatory power. This is partially the result of omitting the unavailable education and vehicle ownership measures, and partially the result of utilizing disaggregate coefficients of disaggregate measures in conjunction with redefined aggregate measures (the car ownership and travel time measures).

5.4 MODE CHOICE

Two models were developed of choice between the main groups of modes (car, public transport and slow modes). One model is for work tours, the other for education tours. Choice between main modes for other travel purposes is represented in the joint models described in Section 5.5.

5.4.1 TOPICS IN MODE CHOICE MODELLING

For work and education tours, it is reasonable to assume that the choice of

destination is determined at a level more fundamental than the choice of mode, and that mode choice should therefore be modelled conditionally on destination choice*. This theoretical argument is quite strong, so that although a joint choice of mode and destination would give better estimates of some of level-of-service parameters, the sequential structure was retained for these two purposes.

The problem arising with the level-of-service measures is that there is little variance between the alternatives. For example, the cost of a public transport journey and the cost of the car journey to the same destination may be highly correlated, and estimation of a mode choice model will therefore give little information about the importance of this variable. In less formally statistical terms, cost is not (in present circumstances) an important influence on mode choice: the crucial variables are the accessibility of public transport and the traveller's own preference for one mode or another (represented in tese models by socio-economic segmentation variables). In future circumstances, of course, if costs change then mode choice will change and an appropriate estimate of the sensitivity of mode choice to costs is therefore essential. If it were technically possible to estimate simultaneously the mode choice and destination choice models, without abandoning the conditionality of mode on destination, the better estimates could be achieved. In practice, because sequential estimation had to be applied rather awkward devices had to be adopted in the attempt to obtain good estimates of cost parameters and other parameters related to tour lengths,

One of these devices was the use in the work mode choice model of the information that was available about the exact location of the traveller's destination. This information is available from the survey to the accuracy of 500 metres, just as is the traveller's origin. However, these more prerise destination were not coded into the networks, as this would have been too much work, nor can they be used in models including choice of destination, since these must be at a zonal level for the choice itself. This more exact information would therefore be used only when destination choice is fixed and when crow-fly distance is an acceptable substitute for network measures. For the slow mode in the work mode choice model, these conditions are satisfied, and the improved accuracy of the modelling is very helpful in obtaining good estimated of the journey length parameters. It is necessary, however, in the final version of the model to use only location variables defined at the

---

* This argument appears to apply equally well to both "work" purposes.

zonal level, since logsums must be taken from this model for use in the zonally defined destination choice model (of course in the "zonal" version of the model, both origin and destination are defined zonally).

To achieve both the improved definition possible from the use of 500M-accuracy destinations and the consistency with zonal definition required for consistency with the destination choice model, two models were developed. The zonal-destination mode choice model used much of the information derived from the 500M-destination mode choice model: this is explained in more detail in the following sub-section.

There appeared little difference in mode choice behaviour between travellers with the two different "work" purposes. For this reason, and because the number of observations for those not going to their "fixed workplace" was relatively small, a single model was developed for mode choice for these two purposes.

In the education mode choice model, the problems were slightly different. It is in the nature of education tours that the overwhelming majority are made by children, who cannot drive a car, and that most of the tours are very short. These tours are therefore overwhelmingly made by slow modes: roughly 90 percent. A mode choice model has therefore as its chief aim the identification of the relatively small minority who may use a motorized mode, and the choice between the motorized modes is somewhat secondary. A danger for these models is that, since it is the people who travel furthest to school that use motorized modes, the time and cost variables for those modes are naturally positively correlated with the choice of the modes, and it is difficult to obtain a realistic negative value for the parameters of those variables.

The method that was adopted to obtain the model recommended was to suppress the car sub-mode choice model, representing four alternatives (slow, public transport, car-driver, car passenger) in the main mode choice. This had two benefits for the modelling. First, the parameters for slow and public transport modes could be estimated taking into account the car availability of the traveller (i.e., whether or not he had a licence). Despite car drivers being only 2.4 percent of the travllers, more that half of those who could drive a car did so. Second, the parameters for the car modes could be estimated from a data set of reasonable size. In Section 5.3 it was reported that the number of education travellers' with a car available was insufficient to estimate a reasonable model.

Despite this restructuring which greatly improved the model, it was not possible to obtain independent estimates of time and cost for the motorized modes. A generalized time variable was therefore calculated using an assumed "value of time". It is not felt that this unsatisfactory feature is a serious problem because of the small market of the motorized modes for this purpose.

5.4.2 THE WORK TOURS MODELS

As mentioned in the previous section, it was necessary to develop two models of work mode choice to solve the problem of high correlation among the journey length variables.

The first of these models is shown in Table A4.1. The crucial variables in this model (10-12, and also 26 and 27) use the straight-line distance between origin and destination, measured to the 500M accuracy to which locations were coded. Of course, there is a loss in using straight-line distance rather than the more accurate approach of taking distance from a network, but in this case the loss is more than compensated by the gain in accuracy in specifying the destination to 500M accuracy rather than at a zonal level.

Although this measure was invaluable in improving the quality of the model, it can be seen from the table that the accuracy of estimation of the time and cost parameters (5 and 6) is not very good. Nevertheless, their relative value to each other and to other variables is reasonable and we are inclined to accept the values estimated.

The other level-of-service measures are the logsums (3 and 4) and the headway variables (7 and 8). There is no reason to think the estimates of these variables are other than satisfactory. A further variable that may be a proxy for level-of-service is the Rotterdam transit dummy (variable 9), which may be a correction for the improved service offered in that city by the metro (not modelled separately from other public transport modes).

Four variables (13, 14, 22 and 25) reflect the influence of the time of travel on the mode choice. These all appear to have natural values. The remaining variables are chiefly socio-economic, reflecting preferences for the modes of different groups in the population. None of these variables raise specific problems.

In summary, the model is reasonably satisfactory. However, the low accuracy

and limited detail achieved for the journey length variables suggest that an approach modelling work mode choice with workplace choice might yield better estimates than those obtained here. The model presented is the culmination of a long process of investigation and experiment, and it is not likely that a much better specification with mode choice conditional on destination choice could be found for the data used.

The auxiliary model exploiting the coefficients of the preliminary model is shown in Table A4.2. This model is a slightly aggregated version of the preliminary model, in which the destination is represented at a zonal level. Variables depending on distance are therefore re-estimated, together with the constants (to avoid bias). In addition, a coefficient is estimated for the compound "utility" term comprising those variables carried over from the preliminary model with their coefficients. This model is merely a technical adjustment to the preliminary model, and gives little further insight into choice mechanisms.

The precise variables used in the "utility" function in this model are numbers 1 to 9 and 13 to 25 from the previous model, multiplied by the coefficients estimated in that model. The coefficient of utility is comfortingly close to 1.0. Further, the distance variables are also generally reduced in magnitude, but have a similar relationship to each other. Note, however, that there is a greatly reduced fit in this model: the likelihood is more than 100 worse than in the preliminary model, showing the great loss of information caused by zonal coding of destinations, despite the gain from using networks rather than straight-line distances.

Zonal Model

The zonal model is a further aggregation of the same preliminary model. In this case the origin is also coded to zonal level and many of the socio-economic variables are omitted, as can be seen from the specification of the model in Table B4.2. In this case the fit to the data is reduced, and the utility coefficient is far from 1.

To compensate for this reduced fit, a few new variables were introduced into the zonal model. These variables normally have an approximate correspondence with a variable that appeared in the original model, but was not present in the aggregate data. Naturally, these substitute variables are less precise than the originals whose place they are taking.

5.4.3 THE EDUCATION TOURS MODELS

The disaggregate mode choice model of education tours, shown in Table A4.3, contains a slow mode sub-mode choice logsum variable, a generalized time variable (constructed with the relative values of walk time, in-vehicle time and cost fixed as input values), and two transit headway measures. Car ownership related to both driving licences and total household size appears in the specification. Age, income, educational status, household size, and occupation are the socio-economic variables used to define market segments amont the trip makers. As in previously reported models, a piecewise linear function of slow-mode distance is included, as is a Rotterdam public transport level-of-service indicator. Almost all coefficients are very accurately measured, and the overall model summary statistics are reasonably good.

The Zonal Model

The specification of the zonal education mode split model (Table B4.3)is very similar to the disaggregate education model. Furthermore, the results are also similar. Again, most coefficients are very accurately estimated, but the overall fit is slightly inferior because of the lower accuracy of aggregate data used for estimation.

5.5 JOINT MODE-DESTINATION CHOICE

The following sub-sections describe the results obtained from the five joint mode/destination models. Joint models were developed for purposes other than work and education because there is little a priori reason in these cases to prefer a mode-destination structure to a destination-mode structure. Further, preliminary tests with a shopping mode choice model indicated that the problems in estimating travel time and cost coefficients found with the work mode choice models (which were estimated a little earlier during the project) would be at least as bad in this case. The most relevant structural issues are discussed in some detail in Section 4.1.

5.5.1 TOPICS IN JOINT MODELLING

For the sub-mode choices four separate purposes were recognized in this group: shopping/personal business social, recreation, and miscellaneous tours. Prelinary analysis for the joint models, however, indicated that the "daily needs" shopping tours were rather different from the durable shopping and

personal business tours. This can be seen in the tables of Report 5, where the tour lengths are particularly different (and with them the mode choice); further differences appear in the amount of time spent at the destination and the frequency of detours to secondary destinations. Accordingly "daily needs" shopping was analyzed separately from the other tours of the group: the two new groups are called "shopping" and "personal business" for simplicity. Other stratification of the tours appeared undesirable on behavioral grounds or impossible because of a shortage of data.

The structuring of mode choice meant that 3 modes (car, public transport and slow) had to be considered in these models, and the decision (explained in Section 4.4) to sample 20 destinations independently of the chosen destination, implied that a maximum of 63 combined alternatives were available to travellers. Because both mode and destination characteristics were included in the models, about 20 variables for each alternative were typically stored on the files used for estimation. These large sizes imply the storage of about 1200 data items per traveller, and to make estimation runs feasible on the computer it was necessary to reduce the number of travellers considered on a given run significantly below 1000. Uniform sampling (e.g.,every other tour) would have been simple, but would have made public transport tours very infrequent, so a sampling by mode was introduced (also outlined in Section 4.4). The results of this sampling are shown in Table 10 .

In all of these models attraction or "size" variables appear. Estimation using such variables, particularly when more than one appears in a model, is subject to certain constraints and limitations. These are explained in Section 4.3

5.5.2 THE SHOPPING TOURS MODELS

Table A5.2 shows the results of the joint mode/destination model of shopping tours. Note that the last two variables, CTS-POP(J) and CTS-RET(J) are "size" (or "attraction" variables), and that no coefficient has accordingly been estimated for the latter.

Two variables capture systematic taste variation as a function of socio-economic characteristics: C-F-LD-CO (female, licenced driver, car owner), showing that an individual in this category is more likely to travel for shopping by car regardless of the destination; and C-CARS/LIC (cars per driving licence), showing that decreasing competition for cars or increasing car ownership leads to increasing car usage. "Environmental" influences are measured by the vari-

ables (1) S-HAAG-tours with destinations in The Hague are less likely to use slow modes for shopping (this result is contrary to the previously reported model of mode choice for work tours); (2) T-RDAM-tours with destinations in Rotterdam are more likely to use public transport (in accord with previously reported results); (3) C-CBD-ORG-tours with origins in downtown areas are less likely to use car, probably because of perceived parking difficulties on returning home; and (4) CTS-CBD-DES-shopping tours are more likely to be made to central zones, probably because of the many alternative shopping opportunities found in those places.

Table 10. Data Sets Used in Joint Model Estimation Page 1 of 1

	SHOPPING	PERSONAL BUSINESS	SOCIAL	RECREATION	MISCELLANEOUS
Tours Total	1454	898	1205	1032	1408
Sample Rate					
Car	1	1	1	1	1
Public Transport	1	1	1	1	1
Slow	1/4	1/4	1/2	1/2	1/2
Other	0	0	0	0	0
Tours Sampled	561	480	838	652	943
Tours Used (1)	447	352	680	552	753
(2)	517	437	731	613	874

(1) In fully disaggregate model.
(2) In zonal model (filters for missing data are less severe).

The other variables in the model describe various characteristics of alternatives (some mode specific but to all destinations, some destination-specific but for all modes, and some specific to particular mode-destination pairs). In this category falls the car logsum variable, measures of travel time (in-vehicle time, walk time, and headway as an indicator of wait time), cost, distance, and densities and sizes of possible destinations. These variables are to be expected in such a model, and the estimated parameters have intuitively reasonable sizes and magnitudes.

The Zonal Model

The zonal version of the shopping model for mode/destination choices is dis-

played in Table A5.2. The two versions are very similar, as indeed are the overall levels of their summary statistics. The logsum variables have been onstrained to 1.0 in the aggregate model because the unconstrained value was slightly over one (theoretically unacceptable). Four other variables have been omitted from the aggregate specification, one has been modified, and two have been added. The omitted variables are T-RDAM (Rotterdam), S-DISR8(J) (part of the slow-distance function), C-F-LD-CO (female licensed-driver,car-owner), and CTS-RET-DEH (J) (retail employment density). They were deleted because of the limitations of the aggregate data set or because the estimated parameters were not reliable. The modified variable was cars per licenced driver which was changed to a car ownership indicator because of data set limitations. The new variables are the intrazonal indicators for car and slow modes, which arise because no measures of intrazonal tour characteristics are available from the zonal network. Other variables, their coefficients, and the coefficient estimation accuracy are all quite similar.

5.5.3 THE PERSONAL BUSINESS TOURS MODELS

The specification of the joint mode/destination choice model for personal business tours (see Table A5.3) is very similar to the corresponding shopping model presented in Table A5.2. Indeed, six of the 21 shopping variables have been dropped (distance over 8 kilometers for slow, CBD origin for car, distance less than 2 kilometers for car, distance for all modes, and retail employment and popolation densities), and four have been added (age less than 15 years for car, distance if the tour was late for all modes, intra-zonal tour indicator for all modes, and service employment). While the estimated coefficients of the remaining 15 variables common to both models are sufficiently different to justify separate models, the general pattern of the impacts they measure are quite similar to that of the shopping model. The six omitted variables were dropped primarily because their coefficients were not significantly different from zero or the poor accuracy of their estimation. The new variables reflect a variety of effects: children are likely to accompany parents in car travel for personal business, within-zone travel is more frequent for personal business trips, and the service employment of a zone is a measure of the personal business opportunities it presents,although not as important a measure as retail employment. The overall model fit is comparable.

The Zonal Model

The aggregate model (Table B5.3) has a number of important revisions to the specification of the disaggregate model. Most important is that (1) the level-

of-service variables have been combined into a "utility variable, whose coefficient was quite accurately estimated at 0.38 (such coefficients should theoretically fall within the range of 0.0-1.0); and (2) the logsum variables coefficient has been constrained to 1.0. The zonal model variables 4.12 correspond closely to the disaggregate variables 8-16, and their coefficient estimates are also quite close to one another. The zonal model has three new variables added (variables 13-15: intrazonal indication for car, and a piecewise-linear distance function for all modes with two parameters). These variables supplement intrazonal and distance effects already captured by several other variables in the specification.

5.5.4 THE SOCIAL TOURS MODELS

Table A5.4 shows the joint mode/destination choice model for shopping tours. Almost all the model coefficients are quite accurately estimated and have signs which correctly reflect the hypotheses underlying their inclusion in the specification. The logsum coefficients are properly bounded, level-of-service variables have negative signs, car ownership, availability, and lack of intra-household competition for a car are positively related to car usage, public transport use is associated with age, lateness of travel is negatively related to slow mode usage (slow modes may be perceived as unsuitable for night time travel) and positively related to car usage (a faster mode is important with less time left in the day). Females are less likely to use cars, and late trips are more likely to be made to close-by destinations (again, probably because less travel time is available late in the day).

The Zonal Model

The aggregate version of the joint mode/destination choice model for social tours is shown on Table B5.4. The model is very similar to its disaggregate version-two variables have been omitted, one altered, and three added. The slow mode logsum was dropped because there is no zonal version of the slow-mode choice model. The female indicator variable was omitted to simplify zonal forecasting with the zonal version of the model. Cars per licensed driver was changed to a car ownership indicator to conform with the availability of data in the zonal data set. The new variables are (1) an intrazonal indicator specific to the car mode; (2) an indicator of elderly travellers specific to the car mode (given that a traveller is over 65 years old, he is most likely to take car, second most likely to travel by public transport, and least likely to travel by slow mode); and (3) a second parameter for the re-

lationship between slow mode utility and distance (for the separation of distance-utility function into two piecewise linear segments). Other variables, their coefficients, and the accuracy of their coefficient estimates are all quite similar to the disaggregate version.

Interestingly, the aggregate model summary statistics show a slightly better overall performance than the disaggregate version. The indicates that the additional information afforded by the three new variables more than compensates for the loss of information attended by the aggregate level of detail of many of the other variables, notably the network-oriented level-of-service measures.

5.5.5 THE RECREATION TOURS MODELS

The joint mode/destination choice model for recreation tours is presented on Table A5.5. Virtually all of the variables in its specification are familiar from previously reported models. Once again, the accuracy of coefficient estimates are strong and the signs of the coefficients confirm the effects that were hypothesized.

However, the positive sign on the coefficient of C-SHORT3(J) is somewhat unexpected, since it seems to imply that short trips are more likely to be made by car than by other modes (even though it might be expected that short trips would be made by slow modes, while cars would be used for long trips). In reality, the positive sign of this coefficient is more likely indicating that, all other things held equal, short trips are more likely to be made than long trips. If a new variable was entered into the specification, S-SHORT3(J), identical in all respects except specific to the slow modes, we would expect a positive coefficient, too, with a greater magnitude than the car-specific short-trip coefficient.

The Zonal Model

The aggregate version of this joint recreation tours model is shown in Table B5.5. Again, the zonal version mirrors closely its corresponding disaggregate specification. The lack of a zonal slow sub-mode choice model precludes the use of a slow mode logsum variable; lack of certain zonal data necessitates the use of a car ownership indicator in place of cars per licensed driver and the use of average activity time at the destination in place of actual value; finally, two intrazonal indicator variables were added. The overall statistical fit of the zonal model was only slightly lower than that of the

disaggregate version, and the accuracy of the coefficient estimates were all quite good. Again, the same set of effects are captured in both the disaggregate and zonal versions of the model.

5.5.6 THE MISCELLANEOUS TOURS MODELS

Table A5.6 shows the joint mode/destination choice model for miscellaneous tours. This model contains two variables specific to the "serve passenger" tour purpose, one of the major components of the general category of miscellaneous tours (along with medical, church, and "other" tours). One is a purpose indicator variable specific to the car mode (C-SERPAX), while the other measures the interaction effects between the serve passenger purpose and distance, for all modes (CTS-DISTSP(J)). From these variables we note that serve passenger travel is more likely to utilize the car mode* and more likely to be short than other miscellaneous travel. Also, note that young children are more likely to be car users (i.e., car passengers) compared to older travellers for these miscellaneous tours. The car logsum variables coefficient was constrained to equal 1.0 to conform with the theoretical requirements for such coefficients. The other variables and their resulting coefficient estimates are reasonably similar to previously examined models.

The Zonal Model

The zonal version of the miscellaneous tours model for mode/destination choice is shown on Table B5.6. The typical changes to convert a disaggregate specification to a zonal one was implemented in this instance. As is usual, the slow-mode sub-mode choice logsum variable was omitted due to the lack of a zonal version of such a sub-mode choice model. Cars per licensed driver was converted to a car ownership indicator; similarly, household size was converted to (zonal) average household size. Finally, in-vehicle time (to which was added half the headways, for public transport) and travel cost were combined into a generalized cost variable, whose weights correspond to the disaggregate model's estimated coefficients for time and cost. The model's overall fit is very nearly as good as that of its corresponding disaggregate verison.

5.6 DESTINATION CHOICE MODELS

As mentioned in Section 5.4, mode choice and destination choice were not mo-

---

* Serve passenger tours were also observed for slow modes.

delled jointly for work and education choice. The mode choice models for these purposes were described in 5.4, and the current section describes the destination choice models.

5.6.1 TOPICS IN DESTINATION CHOICE MODELLING

For work and education tours, the concept of "choice" of destination is, of course, somewhat inaccurate. For these tours, the destination is largely fixed by the labor market, the employer's requirements or the education authorities and the individual's preferences have only marginal impact. For forecasting purposes it is necessary, however, to predict the outcome of these processes, and that is the objective of the models presented in this section.

The three processes mentioned in the previous paragraph are extremely different in type, and it is necessary therefore to develop three separate models: for work tours to "fixed" or usual workplace; for other work tours; and for education tours.

For work tours to usual workplaces, perhaps the most important consideration is the constraint of the labor market: clearly no more people can work in a given zone than there are jobs in that zone. Fortunately, data giving the number of jobs in each zone of the study area was available and could be used to achieve the "balancing" which is a familiar feature of modelling travel to work. Typically, a balancing factor is calculated for every zone in a study area, but is not reasonable to estimate this many parameters. Accordingly, the models presented here incorporate approximate balancing factors, usually based on aggregations of the study area zones. This procedure avoids the most pronounced biases that would result from the omission of balancing factors.

Some data were available describing the number of workplaces per zone for different classification of workers. Of course this data can give much improved information about the real possibilities of alternative jobs, than the simple total number of jobs in a zone. The value of such data is limited by a number of factors for a given person, however. First, the choice of job for a worker is more determined by his or her particular skill (e.g., secretary, carpenter) than by the industry in which he or she might be working (secretary and carpenter could both work (say) in the docks); the available data, however, tend to give employment by the type of industry. Second, the

balancing process would become extremely complicated if an attempt was made to segregate different groups of workers. Finally, the burden of forecasting employment by groups was felt to be quite severe. Despite these arguments, it did appear advantageous to separate farmers from other workers, because their workplaces are so extremely differently distributed, and this distinction appears in some of the models.

In representing the resistance of workers to travel long distances to their work, industrial classifications were found to be of little value. Socio-economic variables, particularly sex and education, were influential. Further investigation showed that perhaps the key variables was the amount of work that was done. If is natural to expect that part-time workers will travel shorter distances, and it is also reasonable (after a little thought) that those who work very long hours will want to work near home. These effects are to be found in the models.

A fundamental question raised with this analysis is that the direction of causality is unclear. To avoid added complexity, it was necessary to make the usual assumption that the workplace is chosen conditional on the home. This is, of course, a simplification and a study concentrating on longer-term population movements would require a more sophisticated treatment. For a traffic study, it is inevitable that some simplifications must be made, but it is not felt that this particular simplification could have a large impact on the forecasting accuracy of the models.

Balancing factors generally apply to all tours to a destination irrespective of the origin of those tours. It is also possibles to imagine effects that are specific to both origin and destination. An important example in the study area is Zoetermeer, where the recent increase in the population is largely made up of people moving from The Hague. It is reasonable to expect that those people will continue for some years to hold jobs in The Hague, and that there will be more work tours from Zoetermeer to The Hague than could otherwise be explained. Variables were therefore included in the models to represent a number of such effects that were expected to be important: several were found to be significant. It cannot be expected that these parameters will remain constant over time, and some caution in exercising the model for forecasting is therefore essential.

For work tours other than to fixed workplaces, it is difficult to argue that balancing factors should be included and such factors were omitted from the