

---

# **INSTITUTE FOR TRANSPORT STUDIES**

## **Updating Car Ownership Forecasts**

### **Final Report to:**

**The Department of the Environment, Transport and the Regions**

---

Title: Updating Car Ownership Forecasts  
Compilers: Gerard Whelan, Ken Fox, Andrew Daly  
Reference Number:  
Version Number: 1.0  
Date: 4/09/2001  
Distribution: Restricted  
File: D:\NRTF2000\FINALREPORT.DOC  
Authorised by:

Signature:

# CONTENTS

<b>1 SECTION ONE .....</b>	<b>3</b>
1.1 INTRODUCTION .....	4
1.2 EXISTING METHODOLOGY – NRTF 1997 .....	4
<b>2 SECTION TWO – MODEL CALIBRATION .....</b>	<b>6</b>
2.1 INTRODUCTION .....	7
2.2 IMPACT OF COMPANY CARS .....	7
2.3 SATURATION LEVELS .....	9
2.3.1 <i>Variation of Saturation Levels</i> .....	9
2.3.2 <i>The Direct Estimation of Saturation Levels - Methodology</i> .....	12
2.3.3 <i>The Direct Estimation of Saturation Levels – Practical Estimation</i> .....	13
2.4 GROWTH IN LONDON .....	19
2.5 AGE AND LICENCE-HOLDING .....	19
2.6 MISCELLANEOUS CONCERNS .....	20
2.7 EMPLOYMENT AND INCOME .....	20
2.8 MULTI-VEHICLE HOUSEHOLDS .....	20
2.9 INSTITUTIONAL OWNERSHIP .....	21
2.10 SENSITIVITY TO OWNERSHIP AND USE COSTS .....	21
2.10.1 <i>Incorporating Ownership Costs</i> .....	21
2.10.2 <i>Incorporating Running Costs</i> .....	21
2.10.3 <i>Calibration</i> .....	22
2.11 MODEL SYNTHESIS .....	22
2.12 RECOMMENDED MODELS FOR FORECASTING .....	26
<b>3 SECTION THREE - FORECASTING .....</b>	<b>27</b>
3.1 INTRODUCTION .....	28
3.2 BASIS OF THE PROCEDURE .....	28
3.2.1 <i>Disaggregate Models and Sample Enumeration</i> .....	28
3.2.2 <i>Prototypical Sampling</i> .....	29
3.2.3 <i>Optimisation</i> .....	29
3.2.4 <i>Discussion of Method</i> .....	31
3.3 PRACTICAL CONSIDERATIONS .....	32
3.3.1 <i>Definition of Targets</i> .....	32
3.3.2 <i>Correcting Total Population</i> .....	32
3.3.3 <i>Control for Income</i> .....	33
3.3.4 <i>Introducing Company Cars</i> .....	34
3.4 SPECIFICATION OF CATEGORIES .....	34
<b>4 SECTION FOUR - MODEL PERFORMANCE.....</b>	<b>36</b>
<b>4.....</b>	<b>37</b>
4.1 PERFORMANCE.....	37
4.2 SAMPLE FORECASTS .....	38
4.3 CONCLUSIONS.....	39

# **1 SECTION ONE**

## 1.1 Introduction

The work reported here has been conducted on behalf of the Department of Transport, Local Government and the Regions. It aims to address the research issues set out in the Department's invitation to tender entitled "Updating Car Ownership Forecasts". The work centres on two key areas. In the first instance, the structure of the existing national ownership model is improved to incorporate a range of policy sensitive variables, and the model calibration process updated to take account of new estimation techniques. Secondly, a new forecasting methodology is introduced to generate local forecasts of car ownership across all regions in Great Britain. This work forms part of an ongoing process of incremental change to the National Transport Model.

To place this work in context, an outline of the car ownership forecasting methodology used to produce the National Road Traffic Forecasts-1997 is given in the remainder of this Section. In Section 2, we provide a description of the work undertaken to enhance the models. In Section 3, we introduce a new forecasting procedure known as prototypical sample enumeration. Finally in Section 4 we provide the results of some initial testing of the model, draw our conclusions and make recommendations for further work.

## 1.2 Existing Methodology – NRTF 1997

This existing ownership model uses information on household income, household-type (defined by the number and age structure of residents), car-type (defined as the number of vehicles a household chooses to own) and area-type (loosely defined by population density) to derive a probability that a given household will own 0, 1 or 2+ vehicles.

Two separate models are calibrated on pooled cross sectional Family Expenditure Survey data at the household level. The first shows the probability that a household will own at least one vehicle ( $P_{1+}$ ) and the second shows the conditional probability that a household will own two or more vehicles ( $P_{2+|1+}$ ).

$$P_{1+} = \frac{S_{1h}}{[1 + \exp(-LP_{one})]} \quad (1)$$

$$P_{2+|1+} = \frac{S_{2h}}{[1 + \exp(-LP_{two})]} \quad (2)$$

and

$$P_{2+} = (P_{1+}) \cdot (P_{2+|1+}) \quad (3)$$

$$P_1 = (P_{1+}) - (P_{2+}) \quad (4)$$

$$P_0 = 1 - P_{1+} \quad (5)$$

S is the saturation level (an assumed maximum number of cars per household) and LP is termed the 'linear predictor'. This predictor is the linear combination of explanatory variables used during the estimation process.

Both household car ownership models were calibrated using standard logistic regression techniques involving the multi-stage process shown below:

$$LP_1 = d + d_t + (f + b_h)\log(G_t Y) \quad (6)$$

$$LP_2 = LP_1 + d' + (b_a + f')\log(G_t Y) \quad (7)$$

$$LP_3 = k_t + (b + b_h + b_a)\log(G_t Y) \quad (8)$$

$$LP = k + k'(LPA_t) + (b + b_h + b_a)\log(G_t e_a Y) \quad (9)$$

In the first instance  $LP_1$  is taken as a function of a constant ( $d$ ), a year specific constant ( $d_t$ ) and household income ( $Y$ ) segregated by household category ( $h$ ).  $G_t$  is the annual growth over the base period in income in period  $t$ .

Next, the 'residuals' of the  $LP_1$  model are explained in  $LP_2$  as a function of a new constant ( $d'$ ) and household income segregated by area-type ( $a$ ).

Models  $LP_1$  and  $LP_2$  are then combined (without re-calibration) to form  $LP_3$ , where  $b=f + f'$  and  $k_t=d + d_t + d'$ .  $LP_3$  was not estimated directly over the period 1971 to 1991 since area type factors were unavailable in 1976 and 1981.

Further modifications are made to the model to derive  $LP$ . Here the constant  $k_t$  in  $LP_3$  is adjusted by the number of licences per adult to  $k + k'(LPA_t)$ .

In order to replicate the base market shares in different areas, base income ( $Y$ ) in the final model ( $LP$ ) is adjusted by  $e_a$  in each area-type until base market shares are correct.

Following a review of the NATCOP models carried out by the University of Leeds (Whelan, 1999) a new functional form for the linear predictor was specified. This functional form could be estimated simultaneously rather than by the multi-stage process outlined above.

$$LP = k + k'+t(LPA) + (b + b'+b_h + b_a)\log(Y) \quad (10)$$

Where:

$k$  is a constant

$k'$  revised constant for 1976 and 1981.

$t$  is a time trend coefficient that is attached to  $LPA$

$LPA$  is licenses-per-adult

$b$  is the coefficient on the log of income term

$b'$  is a revised coefficient for income for 1976 and 1981

$b_h$  a modifying parameter based on household category

$b_a$  a modifying parameter based on area type

$Y$  income (this has not been adjusted to take account of differences between regions)

Note that  $k'$  and  $b'$  relate to the years 1976 and 1981. They are included to capture the impact of omitting area type impacts in those years and are not required for forecasting. In all instances they were statistically insignificant and are not reported in the results.

## **2 SECTION TWO – MODEL CALIBRATION**

## 2.1 Introduction

A review of the existing NATCOP ownership models highlighted a number of areas in which it was thought that the models could be improved. In particular, additional consideration should be given to the research issues set out in the invitation to tender, including:

- The Impact of Company Cars;
- Ownership Saturation Levels;
- Growth in London;
- Age and Licence-holding;
- Employment and Income;
- Multi-vehicle Households;
- Institutional Ownership; and
- Sensitivity to Ownership and Use Costs

Each of the research issues is addressed in turn before combining the findings to develop a new set of final models.

The dataset on which the models were calibrated and any assumptions made are documented in Appendix 1.

## 2.2 Impact of company cars

In the past it was suggested that the provision of company cars had relatively little impact on 1+ car ownership because it was said “the household involved would have at least one car anyway”. Whilst this is a reasonable assumption to make, omitting reference to company cars within an ownership model ignores the possibility that “households with a company car are more likely to own a second car than are comparable households whose first car is privately purchased” (Tender Document).

Because the FES data set does not contain information on the ownership status of the household vehicle fleet it was necessary extract data from the 1991 National Travel Survey to test this hypothesis. The new ownership models, calibrated on a joint FES/NTS data set have the same structure as the existing NATCOP models but include an additional variable to account for the company car issue. Further modifications to the form of the linear predictor are also needed to take account of differences in the definitions of area type in the two data sets. The modified “linear predictor” is shown below:

$$LP = k_{FES} + k_{NTS} + t(LPA) + (b + b_h + b_{aFES} + b_{aNTS}) \log(Y) + gCC_{NTS} \quad (11)$$

Where:

$k_{FES}$  is a constant where FES data is used

$k_{NTS}$  is a constant where NTS data is used

$t$  is a time trend coefficient that is attached to LPA

LPA is licenses-per-adult

$b$  is the coefficient on the log of income term

$b_h$  a modifying parameter based on household category

$b_{aFES}$  a modifying parameter based on area type definitions in the FES database

$b_{aNTS}$  a modifying parameter based on area type definitions in the NTS database

Y income  
g is the coefficient on the company car dummy  
CC<sub>NTS</sub> is a company car dummy variable.

A company car dummy variable is included in the P<sub>2+1+</sub> model and is set equal to 1 if one of the household's stock of vehicles is provided by a company, else it is 0. This variable is also included in the P<sub>3+|2+|1+</sub> model together with an additional company vehicle dummy which it is set equal to 1 if two of the household's stock of vehicles is provided by a company, else it is 0.

Prior to the calibration of a joint FES/NTS model, both data sets were examined to ensure that they have similar properties and that models calibrated on each data set individually have similar estimated coefficients. Following this examination we concluded that it is acceptable to merge both data sets.

**Table 1: Car Ownership Models Incorporating Company Vehicle Dummies**

	Model 1+	Model 2+ 1+	Model 3+ 2+ 1+
K (ASC)	-19.45 (72.9)	-18.68 (46.9)	-9.733 (11.7)
t (LPA)	2.679 (22.4)	3.990 (22.4)	2.958 (6.5)
b (ln Income)	1.768 (62.9)	1.324 (32.8)	0.5533 (6.5)
HH1 (1 adult, no children)	0.0	0.0	0.0
HH2 (1 adult, retired)	-0.08332 (13.5)	-0.06692 (2.2)	0.0
HH3 (1 adult, with children)	0.04094 (5.7)	0.03141 (1.5)	0.0
HH4 (2 adults retired)	0.03911 (6.8)	0.06688 (4.9)	-0.2193 (2.9)
HH5 (2 adults, no children)	0.05377 (11.7)	0.1592 (16.0)	-0.03708 (1.4)
HH6 (2 adults, with children)	0.07341 (15.9)	0.1806 (18.3)	-0.05353 (2.1)
HH7 (3+ adults, no children)	0.01885 (3.1)	0.2550 (24.5)	0.1180 (4.6)
HH8 (3+ adults, with children)	0.003449 (0.5)	0.2325 (21.5)	0.1043 (4.0)
Area2 (metropolitan districts)	0.02486 (4.4)	0.01609 (2.2)	0.0
Area3 (FES pop. den. 1)	0.04905 (8.9)	0.03898 (5.6)	0.03702 (3.3)
Area4 (FES pop. den. 2)	0.09867 (16.1)	0.07138 (10.2)	0.04892 (4.7)
Area5 (FES pop. den. 3)	0.1217 (18.6)	0.07594 (10.6)	0.04173 (3.9)
Area3 (NTS pop. den. 1)	0.1045 (8.6)	0.04325 (3.4)	0.04935 (2.2)
Area4 (NTS pop. den. 2)	0.1332 (10.8)	0.05947 (5.1)	0.05178 (2.6)
Area5 (NTS pop. den. 3)	0.1577 (12.5)	0.07798 (6.7)	0.003665 (0.2)
1 Company Car	n.a.	1.24 (9.6)	0.4970 (2.4)
2 Company Cars	n.a.	n.a.	1.062 (3.1)
Final Log Likelihood	-20907.5131	-13064.2209	-2987.3585
No. Obs.	46137	28472	7838

With regard to company cars it can clearly be seen from Table 1 that the provision of a company vehicle significantly increases the probability that a household will acquire additional vehicles. Other things equal the provision of one company vehicle increases the conditional probability that a household will own a second vehicle from 25.78% to 54.55% and increases the conditional probability that a household will acquire three or more vehicles from 16.51% to 24.53%. Where households acquire two company vehicles, the probability that they will own three or more vehicles is increased from 16.51% to 36.38%.

## **2.3 Saturation Levels**

There are two separate but related issues surrounding the use of saturation levels within the NATCOP models. The first relates to the variation of saturation levels by area type and the second relates to the estimation of saturation levels. The two issues are dealt with in turn below.

### **2.3.1 Variation of Saturation Levels**

Anecdotal evidence suggests that saturation levels vary between area and household types. In the first instance we explore this issue by plotting cars per household against income within each area and household category and examine levels of ownership amongst the highest income households. This essentially graphical analysis allows us to estimate initial saturation levels for each area and household category for the  $P_{1+}$  and  $P_{2+1+}$  models. These values are then used as starting values for the more sophisticated direct estimation procedures.

(i) Households with 1+ Vehicles

Table 2 shows the percentage of households owning one or more vehicles across different income groups in the FES 1971-1996 dataset. When we examine ownership rates for high-income households, it is clear that almost 100% of households own one or more vehicles. Plots of the ownership rates shown in Table 2 are presented in appendix 2. On the basis of this graphical analysis, we would expect saturation levels in the P1+ model not to be significantly different from 1 for all household and area categories. We should bear in mind that where ownership levels are disaggregated by household, area and income group, the degree of confidence that we have in the estimated ownership rates is reduced. For example, the ownership estimate of 0.83 for income group 7 and household type 3 is based on only 12 observations.

**Table 2: Percentage of Households with 1 or more Vehicles**

	All	HH1	HH2	HH3	HH4	HH5	HH6	HH7	HH8	Area1	Area2	Area3	Area4	Area5
IG1	0.10	0.12	0.05	0.13	0.17	0.25	0.56	na	na	0.10	0.09	0.09	0.13	0.19
IG2	0.34	0.31	0.20	0.21	0.37	0.41	0.46	0.37	0.56	0.28	0.29	0.31	0.40	0.49
IG3	0.59	0.50	0.46	0.45	0.68	0.59	0.64	0.54	0.57	0.42	0.54	0.55	0.69	0.76
IG4	0.74	0.71	0.63	0.64	0.85	0.74	0.78	0.64	0.63	0.61	0.69	0.72	0.86	0.87
IG5	0.83	0.80	0.60	0.78	0.89	0.85	0.86	0.78	0.69	0.73	0.79	0.83	0.90	0.92
IG6	0.89	0.87	0.94	0.79	0.88	0.93	0.93	0.84	0.76	0.77	0.85	0.92	0.95	0.98
IG7	0.91	0.84	na	0.83	0.92	0.93	0.95	0.89	0.85	0.82	0.89	0.92	0.96	0.97
IG8	0.96	0.84	na	Na	1.00	0.97	0.99	0.95	0.92	0.89	0.97	0.98	0.98	0.99
IG9	0.96	0.94	na	Na	1.00	0.97	0.98	0.96	0.93	0.93	0.97	0.97	0.98	0.99
IG10	0.98	1.00	na	Na	na	0.99	0.98	0.98	0.94	0.95	0.99	0.98	0.99	0.98
IG11	0.97	0.80	na	Na	na	0.97	0.99	0.97	0.98	0.86	1.00	0.97	1.00	0.98
IG12	0.98	na	na	Na	na	0.97	1.00	0.97	0.98	0.93	1.00	1.00	0.98	1.00
IG13	0.95	na	na	Na	na	0.93	1.00	0.97	0.91	0.88	0.96	0.97	0.97	1.00
IG14	0.99	na	na	Na	na	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00
IG15	0.98	0.80	na	Na	na	0.99	0.98	1.00	1.00	0.96	0.96	0.97	1.00	1.00

Note : na indicates cells with fewer than 10 observations  
 IG is income group – defined in the Appendix 1  
 HH is Household type – defined in the Appendix 1  
 Area is Area type – defined in the Appendix 1

(ii) Households with 2+|1+ Vehicles

Table 3 shows the percentage of households owning two or more vehicles conditional that they own at least one vehicle across different income groups. Overall the market appears to reach saturation where approximately 80% of households own two or more vehicles but there is considerable variation in maximum ownership levels between area and household categories. With regard to area types it is clear that Area1 (London) has a lower percentage of households with two or more vehicles. This trend should be accounted for in the model. Saturation levels for household types are more difficult to determine. In previous calibrations of the model it was argued that single person households should have saturation levels equal to zero since the ultimate objective of the model was to generate traffic forecasts and it was reasonable to assume that a person cannot drive more than one vehicle at once. If, however, the aim of the model is to forecast vehicle stock then we need to relax this assumption and examine ownership rates for all household types. Graphical plots of the information contained in Table 3 are presented in Appendix 2.

**Table 3: Percentage of Households with 2+|1+ Vehicles**

	All	HH1	HH2	HH3	HH4	HH5	HH6	HH7	HH8	Area1	Area2	Area3	Area4	Area5
IG1	0.13	0.03	0.02	0.05	0.00	0.13	0.15	na	na	0.07	0.05	0.06	0.11	0.06
IG2	0.08	0.04	0.01	0.04	0.01	0.08	0.16	0.10	0.21	0.05	0.06	0.06	0.08	0.08
IG3	0.12	0.05	0.02	0.03	0.07	0.10	0.13	0.19	0.19	0.07	0.07	0.09	0.11	0.17
IG4	0.17	0.06	0.03	0.07	0.08	0.15	0.16	0.30	0.31	0.12	0.13	0.14	0.18	0.25
IG5	0.25	0.07	0.12	0.09	0.16	0.20	0.23	0.39	0.35	0.18	0.18	0.23	0.30	0.32
IG6	0.34	0.07	0.00	0.05	0.20	0.29	0.33	0.45	0.39	0.21	0.29	0.32	0.40	0.41
IG7	0.47	0.04	na	0.20	0.36	0.40	0.47	0.60	0.49	0.33	0.47	0.39	0.54	0.63
IG8	0.56	0.23	na	na	0.39	0.51	0.51	0.71	0.59	0.33	0.57	0.55	0.62	0.64
IG9	0.64	0.00	na	na	0.29	0.59	0.65	0.72	0.65	0.43	0.58	0.61	0.72	0.74
IG10	0.71	0.06	na	na	na	0.69	0.67	0.81	0.68	0.47	0.65	0.68	0.80	0.77
IG11	0.74	na	na	na	na	0.73	0.71	0.87	0.71	0.59	0.75	0.73	0.76	0.83
IG12	0.77	na	na	na	na	0.66	0.80	0.84	0.83	0.57	0.76	0.63	0.88	0.86
IG13	0.77	na	na	na	na	0.69	0.83	0.83	0.80	0.74	0.88	0.61	0.78	0.84
IG14	0.77	na	na	na	na	0.67	0.76	0.88	0.82	0.74	0.75	0.83	0.73	0.75
IG15	0.76	na	na	na	na	0.66	0.80	0.86	0.76	0.57	0.78	0.79	0.82	0.85

Note : na indicates cells with fewer than 10 observations

(iii) Households with 3+|2+|1+ Vehicles

Table 4 shows the percentage of households with three or more vehicles conditional that the have 2 or more vehicles. It can be seen that overall saturation levels are reached at around 33%. With regard to household types it can be seen that single adult households and retired couple households have a very low propensity to own three or more vehicles and households with 3 or more adults have the greatest propensity to own three or more vehicles. As expected London and built-up metropolitan areas have a lower propensity to own three or more vehicles when compared to rural areas. We must bear in mind again that Table 4 and the associated plots in Appendix 2 have been derived from a limited sample and therefore only provide an indication of saturation levels.

**Table 4: Percentage of Households with 3+|2+|1+ Vehicles**

	All	HH1	HH2	HH3	HH4	HH5	HH6	HH7	HH8	Area1	Area2	Area3	Area4	Area5
IG1	0.03	0.00	na	na	na	0.00	0.00	na	na	na	na	na	na	Na
IG2	0.11	0.14	na	na	na	0.06	0.14	na	na	na	0.09	0.07	0.09	0.11
IG3	0.08	0.04	na	na	0.04	0.08	0.08	0.06	0.16	0.15	0.07	0.07	0.14	0.08
IG4	0.11	0.15	na	na	na	0.10	0.08	0.21	0.16	0.07	0.08	0.14	0.15	0.13
IG5	0.10	0.16	na	na	na	0.08	0.06	0.16	0.22	0.11	0.13	0.12	0.11	0.11
IG6	0.12	0.00	na	na	na	0.10	0.07	0.20	0.17	0.09	0.05	0.11	0.16	0.16
IG7	0.14	na	na	na	na	0.05	0.07	0.32	0.23	0.08	0.11	0.15	0.17	0.16
IG8	0.17	na	na	na	na	0.09	0.06	0.35	0.25	0.18	0.16	0.17	0.22	0.20
IG9	0.21	na	na	na	na	0.11	0.07	0.39	0.31	0.16	0.16	0.24	0.19	0.28
IG10	0.22	na	na	na	na	0.14	0.06	0.44	0.34	0.11	0.28	0.31	0.19	0.20
IG11	0.25	na	na	na	na	0.10	0.07	0.52	0.37	0.12	0.18	0.35	0.26	0.23
IG12	0.26	na	na	na	na	0.10	0.07	0.47	0.61	0.00	0.15	0.38	0.31	0.35
IG13	0.26	na	na	na	na	0.14	0.11	0.47	0.56	0.35	0.17	0.26	0.32	0.25
IG14	0.28	na	na	na	na	0.05	0.04	0.52	0.71	0.07	0.50	0.11	0.16	0.50
IG15	0.33	0.33	na	na	na	0.14	0.21	0.57	0.59	0.30	0.26	0.31	0.39	0.38

Note : na indicates cells with fewer than 10 observations

### 2.3.2 The Direct Estimation of Saturation Levels - Methodology

The paper by Daly (1999) shows how to set up a partially-constrained choice model for a binary choice situation, exploiting the ‘tree logit’ structure. It can be extended to allow estimation of the full ‘DOGIT’ model of restricted choice over  $n$  ( $\geq 2$ ) alternatives as originally suggested by Gaudry and Dagenais (1977), as follows.

Suppose that each choice  $a$  has an attractiveness function  $V_a$ , then define a set of  $n^2$  artificial alternatives with attractiveness functions  $V_{ab}$  defined by

$$V_{ab} = V_a + \log \theta_b \quad (12)$$

for a set of positive constants  $\theta$ . Then we define  $n$  composite alternatives, each being a nest containing an original alternative  $b$  and the  $n$  artificial alternatives with the same constant  $\theta_b$ . The composite utility of nest  $b^*$  is then given by

$$\exp V_{b^*} = \exp V_b + \sum_a \exp V_{ab} \quad (13)$$

$$= \exp V_b + \theta_b \sum_a \exp V_a \quad (14)$$

The choice probability of  $b^*$  is given by

$$p_{b^*} = (\exp V_b + \theta_b \sum_a \exp V_a) / \sum_c (\exp V_c + \theta_c \sum_a \exp V_a) \quad (15)$$

$$= (\exp V_b + \theta_b \sum_a \exp V_a) / \{(1 + \sum_c \theta_c) \cdot \sum_a \exp V_a\} \quad (16)$$

The minimum fraction choosing alternative  $b^*$  is then (when  $V_b \rightarrow -\infty$ )

$$p_{\min_b} = \theta_b / (1 + \sum_c \theta_c) \quad (17)$$

and the maximum fraction is (when  $V_b \rightarrow +\infty$ )

$$p_{\max_b} = (1 + \theta_b) / (1 + \sum_c \theta_c) \quad (18)$$

Effectively the parameters  $\theta$  define for each alternative a part of the population that can only choose that alternative; this gives the minimum choice fraction. The maximum choice fraction arises when the alternative is chosen by its ‘captives’ and the rest of the population *that is not captive to other alternatives*. The nest  $b^*$  represents those who *choose* alternative  $b$  and those who are constrained to it (modelled as choosing alternatives  $ab$ ). It is not necessary that every alternative should have a constrained minimum choice fraction, which can be modelled by considering  $\theta_b \rightarrow 0$ .

In the present context, there are only two alternatives and only one of them has a minimum choice fraction. This constrained group are the fraction  $1-S = \log \theta_{\text{nocar}}$  who are constrained *not* to own a car, so that within the alternative no-car a nest structure has to be set up as above to represent the constrained choice. Since there are only two alternatives in the model, it is sufficient to set one of the attractiveness functions, say  $V_{\text{nocar}}$ , to zero and to work with  $V_{\text{car}}$  as the sole function. We can, for example, define

$$V_{\text{car}} = k + k'(LPA_t) + (b + b_h + b_a)(Y) \quad (19)$$

and estimate the parameters along with  $\theta_{\text{nocar}}$  as explained by Daly (1999).

### 2.3.3 The Direct Estimation of Saturation Levels – Practical Estimation

Tables 2 to 4 and Figures 1 to 3 show data from all FES years (1971–1996) with income adjusted to 1996 levels using the RPI index. The graphs therefore only show saturation with respect to income. This graphical analysis is intended to give an indication of likely saturation levels to be used as “starting values” for the more direct estimation procedures in which saturation levels are estimated with respect to the multi-dimensional concept of “utility of ownership” rather than simply income. In this respect the modelled saturation levels are inherent to the model and represent the best statistical fit to the data.

Preliminary examination of the functional form of the Linear Predictor showed that overall model fit can be significantly improved by taking the natural logarithm of income. This logarithmic transformation of income “flattens” the logit function at the upper asymptote thereby reducing the impact of changes in income on car ownership propensity as income increases. At high income levels this has a similar effect on ownership probabilities as imposing a saturation level.

Having examined the functional form of the linear predictor where saturation is constrained to 1, the next stage to the calibration process was to estimate saturation levels using the methodology outlined above. In general, where income was specified as a natural logarithm the estimation procedure found it difficult to identify a saturation level significantly different from one, however where income was not transformed the estimation procedure was able to recover significant saturation levels with values close to what we would anticipate from Figures 1 to 3. One explanation for this occurrence could be that the logarithmic transformation of income takes account of the shape of the ownership curve and implied maximum ownership propensity thereby making the identification of an additional saturation coefficient difficult. When estimating saturation levels directly we include income in absolute terms within the linear predictor.

**(i) Direct Estimation of Saturation Levels  $P_{1+}$**

Table 5 shows coefficient estimates, the implied saturation level and the goodness of fit for three alternate specifications of the  $P_{1+}$  model. In all instances the linear predictor incorporates a constant (k), a time trend (t) and household income - modified to take account of differences between household and area types. Model 1 shows a global saturation level, Model 2 show saturation levels to vary by household type and Model 3 shows saturation to vary by area type. Where saturation levels were not shown to be statistically different from each other, coefficient estimates were constrained to be equal.

**Table 5: P<sub>1+</sub> Models with Varying Saturation Levels**

	Model 1- Global		Model 2- Household		Model 3- Area	
K (ASC)	-3.529 (38.0)		-3.531 (36.5)		-3.546 (37.9)	
t (LPA)	2.431 (17.4)		2.496 (17.3)		2.496 (17.3)	
b (Income)	0.0001341 (24.8)		0.0001709 (20.8)		0.0001525 (20.7)	
HH1	0.0		0.0		0.0	
HH2	-0.00008224 (13.3)		-0.00009699 (13.0)		-0.00008363 (13.4)	
HH3	-0.00002839 (4.5)		-0.00003861 (4.8)		-0.0000378 (4.8)	
HH4	0.0000525 (8.8)		0.00002518 (2.4)		0.0000514 (8.4)	
HH5	0.00003633 (9.1)		-0.000001537 (0.2)		0.00003559 (8.6)	
HH6	0.00005032 (12.6)		0.00001226 (1.7)		0.00004929 (11.8)	
HH7	-0.000005569 (1.2)		-0.00004212 (5.7)		-0.000007648 (1.6)	
HH8	-0.00002119 (4.7)		-0.00005731 (7.8)		-0.000023631 (5.0)	
Area1	0.0		0.0		0.0	
Area2	0.00002107 (5.0)		0.00001719 (3.9)		0.000002017 (0.3)	
Area3	0.00004517 (10.7)		0.00004178 (9.5)		0.0000259 (4.2)	
Area4	0.00009685 (17.2)		0.00009645 (15.9)		0.00007601 (10.6)	
Area5	0.0001227 (17.9)		0.0001278 (16.4)		0.0001005 (12.3)	
Implied Saturation	Global	0.9585	HH1-3	0.8253	Area 1	0.9020
			HH4	0.9264	Area 2-5	0.9650
			HH5-8	0.9637		
Final Log Likelihood	-19648.5653		-19613.0297		-19632.7960	
No Obs	42595		42595		42595	

By including saturation levels within the model the overall fit is substantially improved (the log likelihood of the P<sub>1+</sub> model without saturation is -19876.1882). Because of the way in which this model is implemented it is not possible to test directly whether the saturation level is significantly different from 1, rather an alternative chi-squared likelihood ratio test is required. This test compares the overall level of fit with and without saturation with one fewer degrees of freedom. On the basis of this test the inclusion of saturation levels is very strongly supported by the data.

Further examination of saturation levels by household and area type showed saturation levels to be significantly different from 1 and each other for household category groupings 1-3, 4 and 5-8 and area types 1 and 2-5. On the basis of this analysis a model with varying saturation levels by area and household types was specified – the results of which given in table 6.

The inclusion of saturation levels by household and area type significantly increases the overall level of fit in the model, single adult households and those living in London were found to have low saturation levels and households with three or more adults and those in less densely populated areas have the highest saturation levels. This evidence matches what we believe to be true.

It is however very important to note that a significant improvement in the overall level of fit can be achieved by constraining saturations level to 1 and specifying income in logarithmic form. This transformation increases the log likelihood to -19471.1618.

**Table 6: P<sub>1+</sub> Model with saturation levels by household and area type**

Model 4		Implied Saturation	
K (ASC)	-3.553 (36.6)	Area 1, HH 1-3	0.6720
T (LPA)	2.524 (17.3)	Area 1, HH 4	0.7868
b (Income)	0.0001937 (20.1)	Area 1, HH 5-8	0.9191
HH1	0.0	Area 2-5, HH 1-3	0.8590
HH2	-0.00009960 (13.2)	Area 2-5, HH 4	0.9338
HH3	-0.00004159 (5.2)	Area 2-5, HH 5-8	0.9658
HH4	0.00002317 (2.1)		
HH5	0.00000221 (0.3)		
HH6	0.00001111 (1.5)		
HH7	-0.00004404 (5.8)		
HH8	-0.00005964 (7.9)		
Area1	0.0		
Area2	-0.000006162 (0.9)		
Area3	0.00001812 (2.7)		
Area4	0.00007024 (9.0)		
Area5	0.00009862 (10.7)		
Final Log Likelihood	-19589.2497		
No Obs	42595		

**(ii) Direct Estimation of Saturation Levels P<sub>2+|1+</sub>**

The next stage to the estimation process was to directly estimate saturation levels for different household and area types in the P<sub>2+|1+</sub> model.

Model 5 shows the P<sub>2+|1+</sub> model with a global saturation level equal to 0.8186 which, on the basis of the evidence presented in Figure 2, is quite plausible.

Model 6 shows saturation levels for different household types. The model has been simplified to constrain the saturation levels for single adult household to be the same and to constrain the saturation levels for household types 6 and 7. This was because in earlier calibrations of the model they were not shown to be statistically different from each other at the usual 5% level. The implied saturation levels are generally plausible with the caveat that saturation levels for two adult retired households appears low – this could be a genuine occurrence in that most retired households only require a maximum of one vehicle or it could be due to a cohort effect in that there are few women drivers of retirement age. This “saturation” level may change over time. With regard to the other coefficient in the model, the introduction of household specific saturation level reduces the significance of the household income modifiers.

**Table 7: P<sub>2+|1+</sub> Models with Varying Saturation Levels by Household and Area Type**

	Model 5 - Global		Model 6 - Household		Model 7 – Area	
K (ASC)	-5.229 (37.6)		-5.115 (36.9)		-5.338 (37.4)	
T (LPA)	4.223 (19.8)		4.277 (20.3)		4.304 (19.8)	
b (Income)	-0.000008288 (1.6)		0.000009349 (1.2)		0.000008051 (1.3)	
HH1	0.0		0.0		0.0	
HH2	-0.00008681 (3.6)		-0.00006478 (2.8)		-0.00008528 (3.5)	
HH3	0.000007739 (0.9)		0.00001945 (1.6)		0.00008778 (0.9)	
HH4	0.00002942 (4.8)		0.00006363 (3.8)		0.00002994 (5.0)	
HH5	0.00007051 (16.4)		0.00005158 (6.1)		0.00007087 (17.3)	
HH6	0.00007971 (18.5)		0.00004894 (5.9)		0.00008054 (19.6)	
HH7	0.0001114 (22.5)		0.00007828 (9.1)		0.0001145 (23.2)	
HH8	0.00009895 (19.6)		0.00007504 (7.8)		0.0001028 (20.0)	
Area1	0.0		0.0		0.0	
Area2	0.00001257 (3.8)		0.00001206 (7.8)		-0.000003056 (0.5)	
Area3	0.00002072 (6.4)		0.00002092 (6.6)		0.000005491 (0.9)	
Area4	0.00003699 (11.0)		0.00003690 (11.0)		0.00001312 (2.2)	
Area5	0.00004005 (11.3)		0.00003934 (11.1)		0.00001557 (2.6)	
Implied Saturation	Global	0.8186	HH1-HH3	0.4264	Area1	0.6471
			HH4	0.3032	Area 2 & 3	0.7949
			HH5	0.7491	Area 4 & 5	0.8736
			HH6 & HH7	0.8763		
			HH8	0.8102		
Final Log Likelihood	-11811.8008		-11772.4439		-11792.4579	
No Obs	26007		26007		26007	

Model 7 is specified to give saturation levels for different area types. Earlier specifications of the model showed it to be prudent to combine saturation levels for area types 2 and 3 and area types 3 and 4. As expected Area1 (London) is shown to have the lowest saturation levels followed by metropolitan areas and areas with high population densities, finally low density population areas are shown to have the highest saturation levels. As with the household model, the inclusion of area specific saturation levels reduces the impact of the income modifiers.

Having developed a models showing five significantly different saturation levels for household types and 3 significantly different saturation levels for area type the next stage to the calibration process was estimate different saturation levels for each household type within each area type. This gives a total of 15 different saturation levels. The final model is show in Table 8 below:

**Table 8:  $P_{2+|1+}$  Model with Varying Saturation Levels by Household and Area Type**

Model 7		Implied Saturation	
K (ASC)	-5.152 (36.5)	Area 1, HH 1-3	0.2558 (2.1)
K (Adj 1976/81)	0.3822 (3.1)	Area 1, HH 4	0.4004 (0.7)
t (LPA)	4.337 (20.2)	Area 1, HH 5	0.6053 (2.3)
b (base Income)	0.00003019 (2.3)	Area 1, HH 6&7	0.7272 (5.2)
b (Adj 1976/81)	0.000007767 (1.3)	Area 1, HH 8	0.6265 (1.8)
HH1	0.0	Area 2&3, HH 1-3	0.3409 (1.8)
HH2	0.0	Area 2&3, HH 4	0.2611 (3.5)
HH3	0.0	Area 2&3, HH 5	0.7482 (5.5)
HH4	0.00005677 (2.9)	Area 2&3, HH 6&7	0.8625 (8.1)
HH5	0.00004550 (3.4)	Area 2&3, HH 8	0.7452 (4.8)
HH6	0.00004110 (3.1)	Area 4&5, HH 1-3	0.4183 (0.9)
HH7	0.00007176 (5.4)	Area 4&5, HH 4	0.3272 (2.7)
HH8	0.00007658 (5.3)	Area 4&5, HH 5	0.8017 (7.4)
Area1	0.0	Area 4&5, HH 6&7	0.9242 (9.9)
Area2	0.0	Area 4&5, HH 8	0.8669 (6.3)
Area3	0.00008452 (3.0)		
Area4	0.00001670 (4.5)		
Area5	0.00001866 (4.8)		
Final Log Likelihood	-11756.4808		
No Obs	26007		

The inclusion of saturation levels by household and area type significantly improves the overall level of fit of the model. With regard to the implied levels of saturation it can be seen that, in general, Area1 (London) has the lowest saturation levels followed by Areas 2 and 3 then Areas 4 and 5. This is as we would have expected looking at the data set out in Table 3 and Figure 2. With regard to different household types, it can be seen that with the exception of retired couple households, ownership propensity generally increases with the number of adults in the household. In general, the inclusion of saturation levels reduces the significance of the income modifiers, with the consequence that income modifiers for all single adult households and modifiers for areas 1 and 2 are constrained to equal zero.

### (iii) Direct Estimation of Saturation Levels $P_{3+|2+|1+}$

Examination of Table 4 shows that the  $3+|2+|1+$  market is still emerging and that no strong ownership trends exist. Whilst it is possible to estimate a satisfactory model for this market with saturation levels constrained to equal 1 or a global saturation level is defined, the data does not permit the direct estimation of saturation levels that vary by area or household type.

**Table 9: P<sub>3+|2+|1+</sub> Model with Global Saturation Level**

Model 8	
K (ASC)	-4.092 (11.6)
T (LPA)	4.337 (20.2)
b (Income)	0.000007453 (0.9)
HH1	0.0
HH2	0.0
HH3	0.0
HH4	-0.00006977 (2.4)
HH5	-0.00006906 (0.9)
HH6	-0.00001120 (1.4)
HH7	0.00004508 (5.3)
HH8	0.00003887 (4.6)
Area1	0.0
Area2	0.0
Area3	0.00001101 (3.0)
Area4	0.00001310 (3.7)
Area5	0.00001212 (3.3)
Implied Saturation	0.6648
Final Log Likelihood	-2613.9276
No Obs	7014

Table 9 shows the coefficients and global saturation level for the P<sub>3+|2+|1+</sub> model. Although the saturation level is significant, there are three problems with the estimated coefficients. These problems are corrected in the final calibration of the model.

## 2.4 Growth in London

Analysis of saturation levels in Section 3 indicate that saturation levels in London are significantly lower than elsewhere and that saturation levels may already be reached or approached. Further analysis of NTS data shown in Figure 4, allows an examination of ownership rates per household for inner and outer London. Since 1985 there has been a mildly negative trend (-0.0101 cars per household per year) in ownership levels across the whole of London. With regard to inner London the trend has been mildly positive (0.0108 cars per year) and with regard to outer London the trend has been mildly negative (-0.0146 cars per year). It can be seen from Figure 4 that ownership trends in inner and outer London are broadly similar and therefore can be merged for model calibration.

## 2.5 Age and Licence-holding

The LPA (Licences per Adult) term within the car ownership model is used as a proxy for a time trend. Other things equal, this variable governs the rate at which saturation is approached over time (change in LPA). Initially, the coefficient for LPA was determined outside of the model whereas in subsequent calibrations (Whelan, 1999) the coefficient and its standard error was determined within the model. We agree with the study brief that the rate of change of licence holding may vary by household category and that the impact that this phenomenon has on ownership should be examined further. Although we proposed to develop a separate LPA series for each household structure we were not able to due to difficulties in gathering data. This may not be so important given that the introduction of household and area type saturation levels reduced the significance of the income modifier coefficients. The disaggregation of LPA by household type may therefore be a step too far.

## 2.6 Miscellaneous concerns

Three further issues surrounding the development of car ownership models were identified within the brief; they are (i) employment and income (ii) multi vehicle households and (iii) institutional ownership. Each is issue is dealt with in turn below.

## 2.7 Employment and Income

If a change in the level of household employment has further implications for car ownership other than the financial ones then the omission of an employment variable within the ownership model may lead to biased coefficient estimates and corresponding forecasts. We have calibrated new models including the number of adults in employment. In the P1+ model the effect of the number of workers in the household was non-linear, we have therefore specified two variables - one representing a single household employee and the other representing the number of employees in households with 2 or more employees. In the P2+ and P3+ models, employment effects are linear and have simply been added to the utility function.

**Table 10: Car Ownership Models Incorporating Employment Effects**

	Model 1	Model 2+ 1+	Model 3+ 2+ 1+
K (ASC)	-18.29 (61.9)	-18.44 (43.2)	-9.265 (11.6)
K (Adj 1976/86)	0.1762 (0.4)	0.6283 (0.8)	0.7499 (0.4)
t (LPA)	2.618 (21.5)	3.996 (22.3)	3.044 (6.6)
b (Income Logged)	1.738 (53.3)	1.287 (29.4)	0.4640 (5.0)
b (Adj 1976/86)	0.03392 (0.7)	-0.002992 (0.0)	-0.04767 (0.3)
HH2 (1 adult, no children)	0.0	0.0	0.0
HH2 (1 adult, retired)	-0.07472 (9.6)	-0.04828 (1.6)	0.0
HH3 (1 adult, with children)	-0.04130 (5.5)	0.03191 (1.5)	0.0
HH4 (2 adults retired)	0.05133 (6.8)	0.07085 (4.9)	-0.1740 (2.3)
HH5 (2 adults, no children)	0.05023 (9.9)	0.1421 (13.4)	-0.06969 (2.5)
HH6 (2 adults, with children)	0.06749 (13.4)	0.1627 (15.4)	-0.08369 (2.5)
HH7 (3+ adults, no children)	0.01250 (1.9)	0.2231 (19..4)	0.05182 (1.8)
HH8 (3+ adults, with children)	-0.005672 (0.8)	0.1968 (16.2)	0.02808 (1.0)
Area2 (metropolitan districts)	0.0	0.0	0.0
Area2 (metropolitan districts)	0.02392 (4.0)	0.02042 (2.6)	0.0
Area3 (pop. den. 1)	0.05300 (9.2)	0.04268 (5.8)	0.04064 (3.5)
Area4 (pop. den. 2)	0.1034 (16.3)	0.07546 (10.2)	0.05345 (4.8)
Area5 (pop. den. 3)	0.1269 (18.8)	0.08030 (10.6)	0.04499 (4.0)
1 Adult in Employment	0.1427 (2.8)	NA	NA
2+ Adult in Employment	0.2101 (3.8)	NA	NA
Number of Adults in Employment		0.1626 (6.8)	0.3847 (8.3)
Final Log Likelihood	-20907.5131	-11852.6824	-2598.9473
No. Obs.	46137	26007	7014

## 2.8 Multi-vehicle Households

As households become increasingly wealthy it is likely that they will acquire more cars than there are household members available to drive them. It is foreseeable that vehicles will be dedicated to specific tasks. To take account of increased ownership levels, two existing constraints have been relaxed.

- (i) The model now allows for the possibility for single adult household to own more than one vehicles.
- (ii) The model structure now explicitly takes account of households with 3 or more vehicles.

## 2.9 Institutional Ownership

As noted in Reference C (car ownership work - linkage between projects), the NATCOP model only forecasts car ownership by households and there may be a non-zero number of cars owned by members of the institutional population, for example: military bases, hospitals, nursing homes, etc. together with company fleet vehicle that are not available for employees personal use. At this stage it has not been possible to derive correction factors as we have yet to gain access to the DVLA VID database to assess whether this is feasible.

## 2.10 Sensitivity to Ownership and Use Costs

Because there is limited variation in the data and the six five-yearly estimates of purchase price and running cost are heavily correlated with each other and the time trend (LPA), it was not possible to freely estimate coefficients for a purchase price index and a running cost index. An alternative way of incorporating motoring costs within the linear predictor was therefore sought.

One reasonable approach is to constrain the value of the motoring cost coefficients within the linear predictor so that the associated elasticity estimates match those found elsewhere (see Whelan,1999). This procedure will maintain the statistical validity of the calibrated models though may lead to what is known as an aggregation bias.

### 2.10.1 Incorporating Ownership Costs

The ownership cost elasticity estimates used as a base on which to derive our coefficients for the linear predictor are derived from an aggregate power growth model (reported in Whelan, 1999).

Using Wardman's model we can generate elasticity estimates for each of the six FES datasets. If we assume these elasticity estimates to be "true", the next task is to derive coefficient estimates for the linear predictor. For example, if we assume a linear functional form for the linear predictor, the associated point elasticity for a given year (t) is:

$$\eta_t = \varepsilon_t = d_t (\text{cost}_t) \left(1 - \frac{P_t}{S}\right) \quad (20)$$

Since we know  $\eta_t$  (the "true" elasticity),  $\text{cost}_t$ ,  $P_t$  (the market share of owning a vehicle) and  $S$  (the saturation level), we can infer a value for  $d_t$ . The optimal value for  $d$  over all years is found at a value which minimises the sum of the squared differences between the estimated elasticity  $\varepsilon_t$  and the "true" elasticity  $\eta_t$ .

### 2.10.2 Incorporating Running Costs

Incorporating running costs within the model was achieved in a similar way to incorporating ownership costs. In this instance the DETR supplied a elasticity estimate of -0.1 for the base year. Given information on market shares, running costs and saturation levels it was possible to derive a coefficient estimate for running costs. Since we only have information on the true

elasticity value for the base year we have simply assumed a linear functional form for this variable.

### **2.10.3 Calibration**

Having derived coefficient estimates for ownership and use costs that generate plausible elasticity values the next stage is to re-calibrate the new base model constraining the coefficient estimates of ownership and running costs during calibration.

## **2.11 Model Synthesis**

In the preceding analysis we have addressed each of the methodological issues in turn. In this section we combine our findings to develop a set of final models. In each case we enhance the existing NATCOP model to include:

- sensitivity to ownership and use costs
- sensitivity to changes in the level of company cars
- directly estimated saturation levels that vary by area and household type
- employment effects
- the ability to forecast multi car households (3+) in greater detail

In choosing which model is most appropriate we have to consider a number of features of the model: the level of overall fit, the sign, significance and relative magnitude of the coefficients, and the implied elasticity functions. All models are calibrated on a joint FES-NTS data set and for simplicity of exposition the additional coefficients needed to merge the two data sets, as shown in equations 10 and 11, are not presented as they are not used in forecasting.

Each model is addressed in turn.

Table 11 shows two models, the first is where income is included within the linear predictor in absolute terms and the second is where income is included in logarithmic form. Where income is included in absolute terms we can see that saturation levels are lower in London and for single adult households. The coefficients are generally significant and have plausible magnitudes and the overall level of fit of the model is very respectable for a model of this kind. Although a significant improvement in fit is achieved when income is included in logarithmic form and saturation levels are constrained to equal 1, to be consistent with the other models ( $P_{2+|1+}$ ,  $P_{3+|2+}$ ) we have chosen to use model with income included in absolute terms for forecasting. Figure 5 shows a graphical illustration of the performance of the enhanced P1+ models when compared with the true ownership rates in the data.

**Table 11: Final P<sub>1+</sub> Models**

	INCOME (Absolute)	INCOME (Natural Logarithm)
K (ASC)	-3.118 (32.3)	-17.79 (64.8)
T (LPA)	2.384 (16.6)	2.417 (20.2)
b (Income)	0.0001853 (20.1)	1.749 (59.2)
HH1	0.0	0.0
HH2	-0.00008475 (11.5)	-0.08094 (12.9)
HH3	-0.00003963 (5.2)	-0.04088 (5.7)
HH4	0.00004139 (3.8)	-0.04173 (7.1)
HH5	-0.000001566 (0.2)	0.05221 (11.2)
HH6	0.00001074 (1.5)	0.07117 (15.0)
HH7	-0.00004476 (6.3)	0.01492 (2.3)
HH8	-0.00006268 (8.7)	-0.002369 (0.3)
Area1	0.0	0.0
Area2	-0.000007345 (1.1)	0.02460 (4.3)
Area3	0.0000139 (2.1)	0.04929 (8.9)
Area4	0.00006525 (8.5)	0.09899 (16.2)
Area5	0.00009175 (10.3)	0.1218 (18.7)
Employment	0.1666 (7.0)	0.04027 (2.0)
Purchase Cost	-0.003171	-0.002976
Running Cost	-0.000568	-0.000509
Implied Saturation Area 1, HH 1-3	0.6862	1.0
Implied Saturation Area 1, HH 4	0.7698	1.0
Implied Saturation Area 1, HH 5-8	0.9203	1.0
Implied Saturation Area 2-5, HH 1-3	0.8751	1.0
Implied Saturation Area 2- 5, HH 4	0.9283	1.0
Implied Saturation Area 2-5, HH 5-8	0.9705	1.0
Final Log Likelihood	-21003.9266	-20906.3215
No Obs	46137	46137

Table 12 shows alternative specifications for the final P<sub>2+|1+</sub> model. As with the P<sub>1+</sub> models the coefficient estimates are estimated with a good degree of precision and have plausible magnitudes. Where income is included in absolute terms saturation levels are as we would expect. Area 1 (London) has the lowest saturation levels, and less densely populated areas the highest saturation levels. Relative to other household types, households with one adult or where two adults are retired have the lowest saturation levels.

On the basis of the plot produced in Figure 6 and the value of the final log likelihood that the model with varying saturation levels be used for forecasting.

**Table 12: Final P<sub>2+|1+</sub> Models**

	INCOME (Absolute)	INCOME (Natural Logarithm)
K (ASC)	-1.060 (7.2)	-14.13 (35.3)
T (LPA)	2.125 (10.3)	1.920 (10.7)
b (Income)	0.00004033 (3.8)	1.265 (31.0)
HH1	0.0	0.0
HH2	0.0	-0.05411 (1.8)
HH3	0.0	0.03067 (1.4)
HH4	0.00005869 (6.2)	0.07958 (5.9)
HH5	0.00003019 (2.8)	0.143 (14.7)
HH6	0.00002521 (2.4)	0.1666 (16.6)
HH7	0.00005201 (4.7)	0.2292 (21.3)
HH8	0.00005043 (4.2)	0.2001 (17.2)
Area1	0.0	0.0
Area2	0.0	0.01398 (1.9)
Area3	0.000009363 (3.4)	0.03930 (5.6)
Area4	0.00001669 (4.7)	0.07221 (10.3)
Area5	0.00001827 (4.9)	0.07595 (10.6)
Company Car 1	1.631 (7.9)	1.231 (9.5)
Employment	0.2283 (7.9)	0.1796 (7.8)
Purchase Cost	-0.02498	-0.0234
Running Cost	-0.003926	-0.004004
Implied Saturation Area 1, HH 1-4	0.2932	1.0
Implied Saturation Area 1, HH 5	0.6410	1.0
Implied Saturation Area 1, HH 6-7	0.7599	1.0
Implied Saturation Area 1, HH 8	0.6624	1.0
Implied Saturation Area 2-3, HH 1-4	0.2846	1.0
Implied Saturation Area 2-3, HH 5	0.7482	1.0
Implied Saturation Area 2-3, HH 6-7	0.8632	1.0
Implied Saturation Area 2-3, HH 8	0.7736	1.0
Implied Saturation Area 4-5, HH 1-4	0.3514	1.0
Implied Saturation Area 4-5, HH 5	0.8159	1.0
Implied Saturation Area 4-5, HH 6-7	0.9333	1.0
Implied Saturation Area 4-5, HH 8	0.8887	1.0
Final Log Likelihood	-12918.5972	-13042.8823
No Obs	28472	28472

The final set of models to be addressed is  $P_{3+|2+|1+}$ . Where income is included in absolute terms and a saturation level is freely estimates some problems arise. The saturation level is higher than what we might expect and coefficient estimates for household types 4, 5 and 6 lead to a situation in which increases in income will lead to fewer household owning 3 or more vehicles. For this reason we recommend that saturation levels are set equal to 1 and income included in logarithmic form. Plots of both models are found in Figure 7.

**Table 13: Final  $P_{3+|2+|1+}$  Models**

	INCOME (Absolute)	INCOME (Natural Logarithm)
K (ASC)	-0.5716 (1.7)	-5.013 (5.9)
T (LPA)	0.9601 (1.8)	0.8437 (1.8)
b (Income)	0.0	0.4564 (5.2)
HH1	0.0	0.0
HH2	0.0	0.0
HH3	0.0	0.0
HH4	0.0	-0.1897 (2.5)
HH5	0.0	-0.06684 (2.6)
HH6	0.0	-0.08819 (3.4)
HH7	0.00003761 (11.0)	0.04688 (1.7)
HH8	0.00003002 (8.6)	0.02077 (0.7)
Area1	0.0	0.0
Area2	0.0	0.0
Area3	0.000007677 (2.5)	0.04124 (3.7)
Area4	0.00001025 (3.4)	0.05416 (5.1)
Area5	0.000008236 (2.3)	0.04503 (4.2)
Company Car 1	0.3661 (1.7)	0.4819 (2.4)
Company Car 2	0.9747 (2.6)	1.115 (3.2)
Employment	0.4238 (8.5)	0.4160 (9.3)
Purchase Cost	-0.02517	-0.0234
Running Cost	-0.004283	-0.004004
Implied Saturation	0.7947	1.0
Final Log Likelihood	-2938.5956	-2941.9624
No Obs	7838	7838

## 2.12 Recommended Models for Forecasting

Section 2.11 outlines the final set of models with income specified in logarithmic and absolute terms. On the basis of the statistical properties of the models together with discussion with the DETR we recommend the set of models detailed in Table 14 be used for forecasting. In the preceding analysis income and ownership and use costs were specified in 1996 units whereas the NRTF models are based in 1991 units. Appropriate rescaling has been undertaken to take this into account.

**Table 14: Recommended Models for Forecasting (income in 1991 absolute levels)**

	P1+	P2+ 1+	P3+ 2+ 1+
K (ASC)	-3.110 (32.2)	-1.003 (6.8)	-0.5980 (1.7)
T (LPA)	2.384 (16.6)	2.131 (10.0)	1.020 (1.8)
b (Income)	0.0002118 (20.1)	0.00004616 (3.8)	0.0
HH1	0.0	0.0	0.0
HH2	-0.00009691 (11.5)	0.0	0.0
HH3	-0.00004532 (5.2)	0.0	0.0
HH4	0.00004732 (3.8)	0.00006713 (6.2)	0.0
HH5	-0.000001791 (0.2)	0.00003448 (2.8)	0.0
HH6	0.00001229 (1.5)	0.00002877 (2.3)	0.0
HH7	-0.00005118 (6.3)	0.00005942 (4.7)	0.00004300 (11.0)
HH8	-0.00007167 (8.7)	0.00005764 (4.2)	0.00003433 (8.6)
Area1	0.0	0.0	0.0
Area2	-0.000008401 (1.1)	0.0	0.0
Area3	0.000016 (2.1)	0.00001073 (3.4)	0.000008786 (2.5)
Area4	0.00007461 (8.5)	0.00001910 (4.7)	0.00001173 (3.4)
Area5	0.0001049 (10.3)	0.00002090 (4.9)	0.000009421 (2.3)
Company Car 1	N.A.	1.631 (7.9)	0.3657 (1.7)
Company Car 2	N.A.	N.A.	0.9742 (2.6)
Employment (numbers)	0.1666 (7.0)	0.2283 (7.9)	0.4238 (8.5)
Purchase Cost	-0.003235	-0.02549	-0.02517
Running Cost	-0.000568	-0.003926	-0.004283
Implied Saturation Area 1, HH 1-3	0.6862	N.A.	N.A.
Implied Saturation Area 1, HH 4	0.7698	N.A.	N.A.
Implied Saturation Area 1, HH 5-8	0.9202	N.A.	N.A.
Implied Saturation Area 2-5, HH 1-3	0.8751	N.A.	N.A.
Implied Saturation Area 2- 5, HH 4	0.9283	N.A.	N.A.
Implied Saturation Area 2-5, HH 5-8	0.9705	N.A.	N.A.
Implied Saturation Area 1, HH 1-4	N.A.	0.2929	N.A.
Implied Saturation Area 1, HH 5	N.A.	0.6410	N.A.
Implied Saturation Area 1, HH 6-7	N.A.	0.7601	N.A.
Implied Saturation Area 1, HH 8	N.A.	0.6623	N.A.
Implied Saturation Area 2-3, HH 1-4	N.A.	0.2844	N.A.
Implied Saturation Area 2-3, HH 5	N.A.	0.7480	N.A.
Implied Saturation Area 2-3, HH 6-7	N.A.	0.8632	N.A.
Implied Saturation Area 2-3, HH 8	N.A.	0.7736	N.A.
Implied Saturation Area 4-5, HH 1-4	N.A.	0.3511	N.A.
Implied Saturation Area 4-5, HH 5	N.A.	0.8158	N.A.
Implied Saturation Area 4-5, HH 6-7	N.A.	0.9333	N.A.
Implied Saturation Area 4-5, HH 8	N.A.	0.8885	N.A.
Global Saturation	N.A.	N.A.	0.7946
Final Log Likelihood	-21003.9758	-12919.0592	-2938.6253
No Obs	46137	28472	7838

### **3 SECTION THREE - FORECASTING**

### 3.1 Introduction

The objective of this section is to detail both the methodology and practical application of prototypical sample enumeration techniques to forecasting car ownership at NTEM zone level using the NATCOP car ownership models.

In Section 3.2 we provide an outline of the theory of prototypical sample enumeration, drawing largely from Andrew Daly's PTRC paper (Daly, 1998). This is followed in Section 3.3 by description of the practical considerations surrounding this application of the technique. In addition, we have included a technical appendix that details how the computer program works and how to operate the software.

### 3.2 Basis of the Procedure

The objective of disaggregate modelling as applied to travel demand forecasting is to explain the choices made by individual decision-makers. This approach has proved very successful as a basis for the development of models and through the technique of sample enumeration, disaggregate models have also been used successfully for short-term forecasting. However, because straightforward applications of sample enumeration do not take account of the changing nature of the population (e.g. the general "greying" of the population), longer-term forecasting is not possible. To fill this gap, the technique of *prototypical* sample enumeration has been developed.

#### 3.2.1 Disaggregate Models and Sample Enumeration

A key characteristic of disaggregate modelling is the statistical approach that it inherently takes to the analysis of data. This approach recognises that it is not possible to predict correctly how each individual (or household) in a population will behave, but this does not prevent information being obtained on the variables that *influence* – rather than *determine* – behaviour. The model for each individual is then formulated as

$$\Pr \{ c_i=k \mid K_i, S_i \} = p_k(K_i, S_i) \quad (21)$$

giving the *probability* that the choice  $c_i$  of individual  $i$ , whose characteristics are  $S_i$ , will be alternative  $k$  from the choice set  $K_i$  (which has availability and characteristics specific to individual  $i$ ). It is a primary objective of the modelling then to specify how the alternatives in  $K$  are described and which characteristics  $S$  are relevant. A further important task in the modelling is to determine the form of  $p$  and estimate the values of unknown parameters that appear in it.

In order to make useful forecasts a means must be found to *aggregate*, to derive from a model predicting the behaviour of individuals a forecast of the behaviour of an entire population. An important point is that it is not correct simply to set  $K$  and  $S$  to the average population values and apply equation (21) as if the entire population behaved like a mass of identical average individuals: *this overstates the response to changes*, an effect known as aggregation bias which has long been recognised (e.g. Daly, 1976, Gunn, 1984). Similarly, the model (21) cannot be used directly to calculate elasticities, again this leads to an overstatement of responsiveness.

A technique that does not have this disadvantage is sample enumeration. Essentially, sample enumeration simply applies the model (1) to each member of a sample in turn. Then, *if the*

*sample is representative*, the sum of the forecasts for each individual is the unbiased forecast for the whole population. Formally, the expected demand  $Q_k$  for an alternative  $k$  is given by

$$Q_k = \sum_i w_i \cdot p_k(K_i, S_i) \quad (22)$$

where  $w_i$  is the expansion factor or weight attached to individual  $i$  in the sample in order to make its sum representative of the population. Very often, the sample used for forecasting is the same sample used for model estimation, while the weights  $w$  are determined by the sampling process used.

The advantages of sample enumeration using the basic equation (22) are its simplicity and convenience. The forecasts are unbiased. It is important to note that the procedure of sample enumeration is entirely independent of the form of the model that is used for forecasting: logit, linear, whatever model is used can be applied in this way.

The primary disadvantage of sample enumeration is that a representative sample may not be available, perhaps because the model is being transferred in time or space. In particular this will always be true when a forecast is required over any considerable period, so that a base-year sample can no longer be considered representative.

The conclusion is that the advantages of sample enumeration are substantial in some circumstances and therefore that it would be advantageous to be able to apply the technique more widely. A means was therefore required for generating representative samples for circumstances different in space or time from those for which real samples are available.

### **3.2.2 Prototypical Sampling**

The most obvious way to produce samples representative of future conditions is to generate an artificial population which has, as far as is known, the characteristics of the future population. However, the forecasts that are generally available – e.g. from planning authorities – typically refer to aggregate statistics such as age-sex population distribution, rather than the composition of individual households. A method is therefore required for generating a sample of households that is internally consistent, i.e. that it ‘looks like’ a typical population, while also achieving consistency with such aggregate statistics as are available.

The objective of the method is thus to use an existing household sample to produce a sample that is or will be representative of one or more target areas. The key method used for adjusting the samples is the adjustment of the expansion weights present on the survey records (the FES does not include expansion weights all households are weighted by the total population divided by the sample size). The following section discusses the possible ways in which these weights can be adjusted.

### **3.2.3 Optimisation**

There are two sets of procedures that have been used in practice to produce prototypical samples: Iterative Proportional Fitting (IPF) and Quadratic Optimisation. Both methods rely on the availability of a detailed sample of households that is not directly representative of a specific target area or year. The detailed sample may refer to another area (larger, smaller or elsewhere), another year or both. The objective of the procedure is to create samples that are

representative of target areas, given data for those target areas that is much less detailed in character.

The construction of prototypical samples by the quadratic optimisation method ('QUAD') rests on the recognition that the data for the target area and the base sample may be inconsistent. That is, the method balances the need to meet the target area marginal totals against the wish to retain the detailed relationships between the frequencies of different household types indicated by the base sample. Weights can be given to the relative divergences: in this sense QUAD is a generalisation of the IPF method, which gives exact matches to the marginal totals but sacrifices faithfulness to the original detailed sample.

A further difference between most applications of QUAD and the IPF method explained by Beckman *et al.* is that QUAD constructs its detailed samples by weighting or re-weighting the records of the base sample, rather than by drawing from the base sample with fixed probabilities. This difference has the minor advantage that the rounding errors found in IPF are eliminated, but its more important advantage is that it avoids the additional step of drawing the sample. The output is thus a sample whose size is predetermined and independent of the target area; the fit to the target area is achieved by the weighting.

Re-weighting is applied to all of the households in each of a series of categories, pre-defined to cover the main dimensions of interest for the prediction of travel behaviour. The categories are defined with respect to variables such as household size, numbers of adults, number of workers and the age of the household head.

QUAD is called quadratic optimisation because it can be specified in the form of optimisation with respect to the new frequencies  $\phi_c$  of households of each category  $c$  of a quadratic function for each target area, i.e. (following Daly and Gunn, 1985),

$$\phi = \operatorname{argmin} ( Q ), \quad Q = \sum_t w_t \cdot (z_t - \sum_c \phi_c \cdot x_{tc})^2 + \sum_c (\phi_c - f_c)^2 \quad (23)$$

and

- $w_t$  is the weight attached to the importance of meeting target  $t$ ;
- $z_t$  is the value per household of target statistic  $t$  in the current area;
- $x_{tc}$  is the average amount of target variable  $t$  for a household in category  $c$ ;  
hence  $(\sum_c \phi_c \cdot x_{tc})$  is the predicted total of statistic  $t$ ;
- $f_c$  is the frequency of household category  $c$  in the base sample.

The first term in  $Q$  clearly represents the error in not meeting the target marginal totals for each variable  $z$ , while the second term represents the divergence from the current distribution of households over the categories. The weights  $w$  are introduced so that differential importance can be given to meeting each of the different targets or that the balance between consistency with targets and consistency with base population can be adjusted. In fact, in most applications it has been found satisfactory to set all the  $w$ 's to 1. Setting large values of  $w$  would cause QUAD to find a distribution of households that matched the target totals very well at the expense of substantial departures from the original distribution, i.e. a solution like that given by IPF.

Note that all terms of  $Q$  are on a per-household basis.

The simple form of  $Q$  makes it in principle easy to optimise. Given any starting value of  $\phi$ , the global minimum of  $Q$  is always at the value  $\phi^*$  given by

$$\phi^* = \phi - Q'(\phi) \cdot Q''(\phi)^{-1} \quad (24)$$

where  $Q'$  and  $Q''$  are the first and second derivatives respectively of  $Q$  with respect to  $\phi$ , i.e. Newton's calculation, which converges directly for a function which is exactly quadratic such as  $Q$ . The calculation is particularly easy if the starting value is taken at  $\phi = 0$ .

However, reality requires that constraints be imposed on the values of  $\phi$ , e.g. that  $\phi \geq 0$ , and there is no guarantee that Newton's calculation will give such a result. The procedure that can be used in this case is then an iterative calculation, in four steps as follows.

1. Specify minimum values  $\phi_{\min}$  for  $\phi$  and set  $\phi_0 = \phi_{\min}$  and  $i = 0$ .
2. Perform Newton's calculation as in equation (4) above deriving  $\phi_{i+1} = \phi_i - Q'(\phi_i) \cdot Q''(\phi_i)^{-1}$ .
3. Check whether all free values of  $\phi_i \geq \phi_{\min}$  and that  $Q' \geq 0$  for all constrained values of  $\phi$ ; if so, terminate.
4. Otherwise adjust any  $\phi$  values that are less than  $\phi_{\min}$  to  $\phi_{\min}$ ; free any  $\phi$  values which are constrained and for which  $Q' < 0$ ; set  $i = i+1$  and repeat from Step 2.

This algorithm can be proved to converge to the overall optimum in a finite number of steps, because the set of constraints  $\phi \geq \phi_{\min}$  form a convex set while the function  $Q$  is concave. Each iteration of the algorithm gives a reduction in the value of  $Q$ . This theoretical result is however of limited value, because the number of steps might be quite large. If the number of categories is of the order of 50, as is commonly the case, the maximum number of steps could theoretically be  $2^{50}$ , approximately  $10^{15}$ . In practice, the number of steps turns out to be very limited: typically convergence is achieved in 5 or 6 iterations.

The values  $\phi_{\min}$  can in principle be chosen to be any (non-negative) limits that seem sensible, such as 10% of the frequency of each household category in the base sample. Their function is to prevent unusual, perhaps erroneous, target data from generating an impossible – or nearly impossible – future population distribution.

### 3.2.4 Discussion of Method

The advantage of the method described here is that a close fit to the "targets" is obtained quickly and reliably with minimal departure from the original distribution. While the exact form of the function chosen is clearly arbitrary, its advantage of simplicity seems decisive: there exist no strong theoretical reasons for choosing another form. The use of a formal minimisation has the considerable advantage that the results are reliable and predictable.

A further advantage is the flexibility available to the user to shift the balance between meeting the targets and maintaining the original proportions. It might seem useful, for example, to keep more closely the original proportions in a base year and to give more weight to the targets for a forecast year: this would be achieved by giving higher values to  $w_t$  for the forecast year than for the base year. Similarly, more important targets can be given more weight if required.

### 3.3 Practical Considerations

The preceding discussion is of course entirely theoretical. To put the proposed method into practice requires both the targets and the categories to be defined; certain difficulties arise in each case.

#### 3.3.1 Definition of Targets

For ease, simplicity and consistency with other elements of the NRTF2000 framework we have chosen target variables that can be derived from output from the Scenario Generator for each NTEM zone in each forecast period.

1. Total Population
2. Children (0 to 15)
3. males in full time employment (16 to 64)
4. males in part time employment (16 to 64)
5. male students (16 to 64)
6. males neither employed nor students (16 to 64)
7. males 65+
8. females in full time employment (16 to 64)
9. females in part time employment (16 to 64)
10. female students (16 to 64)
11. females neither employed nor students (16 to 64)
12. females 65+
13. Total Employment
14. Households with 1 person
15. Households with 2 or more people
16. Households with 1 company vehicle
17. Households with 2 company vehicles

#### 3.3.2 Correcting Total Population

The mathematical optimisation above has been formulated on a per-head basis. Partly this is to simplify discussion of the appropriate weights to be applied to the various components of the optimisation, partly it is to deal with a potential problem concerning the total population.

The problem may arise because there is no overall constraint on the frequencies  $\phi$ . In principle, the  $\phi$ 's should add up to the total number of households (of all categories) per head of population, i.e. the inverse of the household size. However, this applies only when the divergence of the sample from the original distribution is zero, or when the divergences exactly cancel out. This cannot be expected to be the case in general, and the total number of households or people in a zone may well diverge from the target number.

It would be possible to incorporate in the quadratic optimisation the further constraint that the  $\phi$ 's should add up to a pre-specified total. However, this would add substantially to the complexity of the program and require further programming and testing work. The following alternative procedure has therefore been chosen.

This alternative is to include among the targets a total of households or population. This target can be given a large weight to ensure that divergences are small. Then we can be sure that the total population will be very close to the target value. It is for this reason that the

number of targets has been increased by 1: a specific target, say the first, contains this total and will be given a very large weight.

The "total" that is recommended for this purpose is the total of people, not of households. The reason for this choice is that it is likely that the quality of data for the number of persons is better than for the number of household.

Weights of 1.0 were used for all targets except the total population, which was given a weight of 10.0. These weightings were based on experimentation to find a satisfactory balance of all the competing objectives.

### **3.3.3 Control for Income**

Clearly the car ownership model system is responsive to differences in income-growth scenarios. The issue is how this dependence should be introduced into the models. Two options are apparently available: to make the growth in income part of the target and expansion factor system; or to apply simple factors to incomes. The latter approach is recommended.

To introduce income growth through a range of targets would introduce the specific complexity in the process that, since (as is currently understood) income information will be available only at the level of the entire country, a two-step generation of the expansion factors (1-zone and 1203-zone) would be necessary instead of the one-step procedure currently proposed. Further, to avoid distortions it would be necessary to introduce some degree of categorisation on income, increasing the number of categories and expansion factors enormously.

The alternative approach, simply to factor all the incomes in the prototypical sample by a suitable factor, is based on the assumption that the current form of the income distribution will remain unchanged, simply that the real levels will change. While this assumption might be questionable, it is doubtful that better information will be available to improve on it.

It should be noted that if employment rates (per household) change in the future, then average incomes will change automatically. The overall correction to incomes must therefore be made after the expansion factors have been calculated.

It is therefore recommended that income growth for future years be modelled by multiplying all incomes by a single factor. This factor should be determined by calculating the total income in the future year after calculating the new expansion factors. A small program will be necessary to achieve this calculation.

It became apparent during the application of the model that the method may not cope well with zones with very high or very low average incomes. For example, if our base sample has an average annual household income of approximately £19,000 and we are forecasting for zone where average household income is £50,000 the forecasts are likely to be biased. To overcome this problem income levels in the base are adjusted for each zone by:

$$\frac{\text{Average household income in zone X}}{\text{Average household income in our sample (after the QUAD re - weighting)}}$$

Where average household income levels in all 1203 zones were generated using the *MapInfo* geographical Information System. A note on the precise methodology is appended.

### 3.3.4 Introducing Company Cars

Company vehicles are introduced into the forecasting procedure using the prototypical sampling method. Because the base FES data does not have information on whether households have company vehicles we have to develop another model to predict this. To this end, we have calibrated a new model (shown in Table 15) on NTS data that gives the probability that a household will own 0, 1 or 2 company vehicles as a function of the household's socioeconomic characteristics. Each household in the base FES dataset is then assigned a probability that they will own 0, 1 or 2 company cars.

**Table 15: Company Car Model**

	Utility – 1 Company Car	Utility – 2 Company Cars
Log (Income 1996)	0.7889 (19.7)	0.2897 (3.2)
Number of Children	0.06436 (3.2)	0.1722 (3.9)
Employment	1.651 (12.7)	2.407 (3.4)
Head of Household (Male)	0.3809 (4.6)	0.6412 (2.0)
Head of Household (Age)	-0.009922 (5.9)	-0.009527 (2.2)
Constant	-11.26 (27.8)	-8.099 (6.8)
Final Log-likelihood	-10660.3400	
Number of Observations	25991	

The model is calibrated on NTS data from 1985 to 1997 and has a multinomial logit structure with three alternatives - 0, 1, or 2 company vehicles. The utility functions for the acquisition of zero company vehicles is set equal to zero and the utility functions for one and two company vehicles are shown in the table above. Please note that in calibration and application, where a household has zero cars then the probability of owning zero company cars is set to one and where a household only has one car, the probability that they acquire two company cars is set to zero. The model therefore simply predicts the ownership status (company, private) of each vehicle in the household.

Next, we need to make assumptions about ownership rates for company vehicles in each NTEM Zone e.g. 10% of households have one company vehicle and 3% of households have two company vehicles, then apply the prototypical sampling strategy. Company vehicles are therefore a target variable the same as any other target variable.

### 3.4 Specification of Categories

The objectives of the classification of the population into categories are as follows:

- the number of categories should be minimised and their definition kept as simple as possible;
- the flexibility of the sample with respect to meeting the targets should be optimised; this means that there should be scope for variation along the dimensions to which the targets relate (e.g. household size, numbers of workers, etc.);
- the coverage of the population should be maximised.

An extensive analysis of the distribution of the population, as revealed in the base sample has been conducted. The key variables considered in this work have been number of adult, number of children, economic activity and age of head of household.

- 1 Not in other categories
- 2 HoH retired, economically active, 0 children, 2 adults
- 3 HoH retired, economically active, 0 children, 3+adults
- 4 HoH retired, economically inactive, 0 children, 1 adult
- 5 HoH retired, economically inactive, 0 children, 2 adults
- 6 HoH retired, economically inactive, 0 children, 3+adults
- 7 HoH not retired, economically active, 0 children, 1 adult
- 8 HoH not retired, economically active, 0 children, 2 adults
- 9 HoH not retired, economically active, 0 children, 3+adults
- 10 HoH not retired, economically active, 1 child, 1 adult
- 11 HoH not retired, economically active, 1 child, 2 adults
- 12 HoH not retired, economically active, 1 child, 3+adults
- 13 HoH not retired, economically active, 2+children, 1 adult
- 14 HoH not retired, economically active, 2+children, 2 adults
- 15 HoH not retired, economically active, 2+children, 3+adults
- 16 HoH not retired, economically inactive, 0 children, 1 adult
- 17 HoH not retired, economically inactive, 0 children, 2 adults
- 18 HoH not retired, economically inactive, 0 children, 3+adults
- 19 HoH not retired, economically inactive, 1 child, 1 adult
- 20 HoH not retired, economically inactive, 1 child, 2 adults
- 21 HoH not retired, economically inactive, 1 child, 3+adults
- 22 HoH not retired, economically inactive, 2+children, 1 adult
- 23 HoH not retired, economically inactive, 2+children, 2 adults
- 24 HoH not retired, economically inactive, 2+children, 3+adults

## **4 SECTION FOUR - MODEL PERFORMANCE**

## 4.1 Performance

Having calibrated a sensible model with plausible properties and proposed a sound methodology for application, the next stage in model development is to examine the model performance when applied to each NTEM zone.

In the first instance we apply each model to each NTEM zone and compare the predicted ownership probabilities with true market shares derived from census information. As expected, when applying of a model calibrated on a random sample of households to diverse set of area types there are some forecasting errors. To overcome these errors we have made some adjustments to the model constants, for example, where the  $P_{1+}$  model over predicts ownership probability for a particular NTEM zone when compared to the true level, we reduce the value of the constant until the true and forecast probabilities are equal. The procedure is undertaken for each model ( $P_{1+}$ ,  $P_{2+|1+}$  and  $P_{3+|2+}$ ) and for each NTEM zone.

Table 16 shows the results of a series of un-weighted OLS regressions of the forecast ownership levels for each NTEM zone against the true levels. Whilst the uncorrected model performs reasonably well in this notoriously difficult forecasting area, the corrected model is able to replicate the base market shares correctly.

**Table 16: Model Performance - regressions**

Model	Uncorrected Model		Corrected Model	
	Slope (t-stat)	Adj R <sup>2</sup>	Slope (t-stat)	Adj R <sup>2</sup>
Zero Cars	1.0326 (168)	0.9459	0.9950 (1859)	0.9987
One Car	1.0045 (349)	0.9822	0.9868 (2102)	0.9987
Two Cars	0.9588 (137)	0.8731	0.9821 (153)	0.9981
Three Plus cars	0.8588 (104)	0.7810	0.9708 (436)	0.9854

Table 17 shows the (weighted) mean cars per household in each area type as shown in the census and as forecast by the uncorrected and corrected models. It can be seen that the uncorrected model slightly over-estimates the number of vehicles in Area 1 and slightly under-estimates ownership in all other areas. To correct for such forecasting errors we have adjusted the constants in each of the three models so that the forecasts match the actual probabilities in each NTEM zone.

**Table 17: Model Performance by Area Type – mean cars per household**

	Actual (1991 Census)	Uncorrected Forecast	Corrected Forecast
Area 1	0.8124	0.8403	0.8139
Area 2	0.7651	0.7825	0.7675
Area 3	0.9414	0.9118	0.9430
Area 4	1.0518	1.0462	1.0536
Area 5	1.2213	1.1101	1.2222
Total	0.9473	0.9216	0.9499

Finally Table 18 shows the actual ownership probabilities together with the corrected and uncorrected forecasts. It can be seen that the uncorrected model marginally over-predicts the percentage of households with zero cars. This problem is corrected by adjusting the model constants.

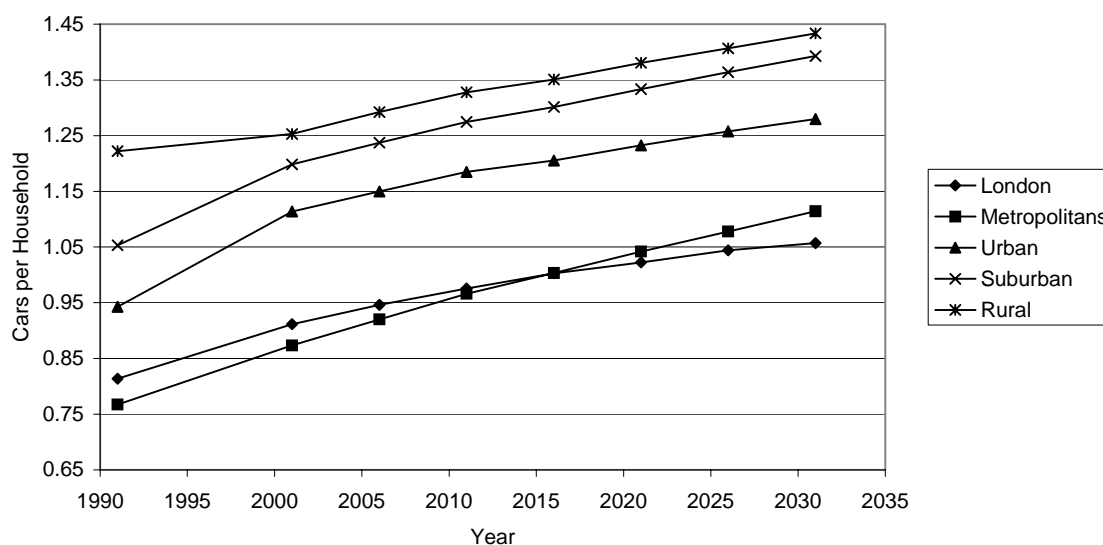
**Table 18: Average Market Shares across all NTEM Zones**

	Actual	Uncorrected Model	Corrected Model
P <sub>1+</sub>	0.6665	0.6422	0.6678
P <sub>2+1+</sub>	0.3471	0.3527	0.3478
P <sub>3+2+1+</sub>	0.1724	0.1886	0.1735
P <sub>0</sub>	33.3	35.7	33.2
P <sub>1</sub>	43.5	41.6	43.6
P <sub>2</sub>	19.2	18.4	19.2
P <sub>3+</sub>	4.0	4.3	4.0

## 4.2 Sample Forecasts

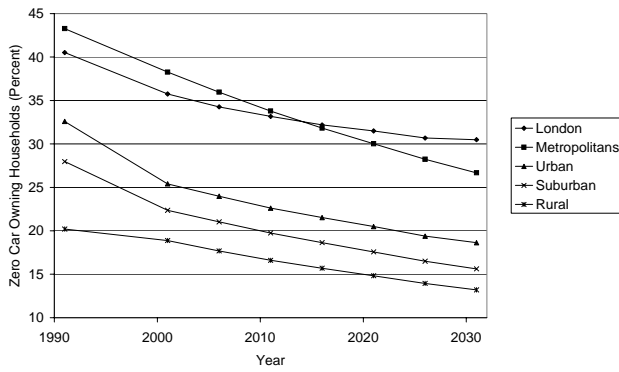
Having validated that the final model can recover actual car ownership levels in each NTEM zone for the base period, the next stage to examine how the model performs over time. To do this we provide an illustrative set of forecasts for the years 1991 to 2031. In this exercise we assume that:

- Income rises at a rate of 2% per annum from the 1991 base,
- Licence holding is fixed at the 1991 level
- Purchase and running costs are fixed at their 1991 level
- Demographics change in line with the Department’s scenario generator.

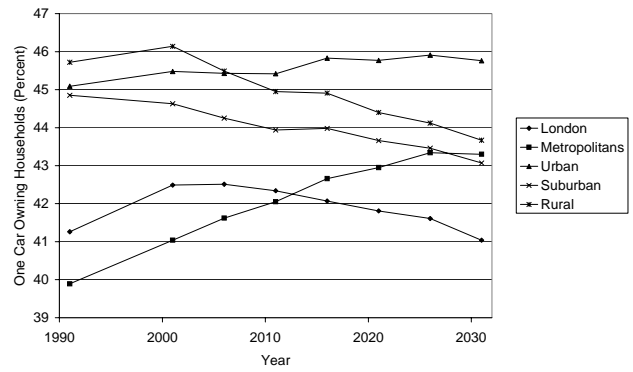


**Figure 1: Forecasts of Cars per Household 1991-2031**

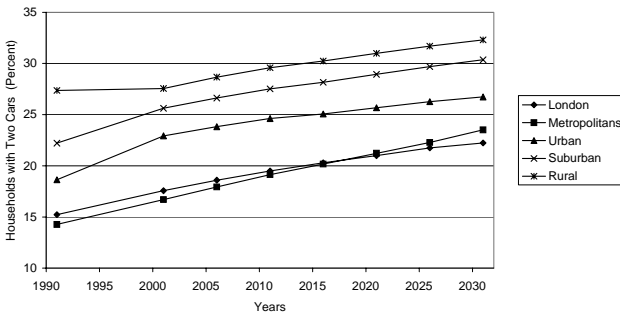
Figure 1 shows forecast of the number of cars per household for five area types over the years 1991 to 2031. Over the forecast period car ownership set to increase by 30% in London, 45% in the Metropolitan districts, 36% in urban areas, 32% in semi-urban/suburban areas and 17% in rural areas. This rate of growth is initially quite strong but diminishes over time as saturation levels begin to bind. A more disaggregate picture of car ownership is given in Figure 2 where the percentage of households owning zero, one, two and three plus vehicles is presented.



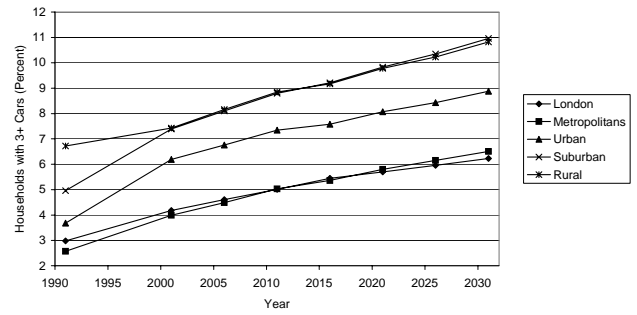
**Figure 2a: Zero Car Households**



**Figure 2b: One Car Households**



**Figure 2c: Two Car Households**



**Figure 2d: Three plus Car Households**

Figure 2 shows that over the forecast period the percentage of households without cars falls, and the percentage households with two or three-plus cars increases. With the exception of the Metropolitan areas, the percentage of one-car households remains fairly stable over the period. This is in line with what we might expect and the patterns between area type conform with the saturation levels estimated during calibration.

### 4.3 Conclusions

In this report we have developed a new set of car ownership models enhanced to incorporate:

- sensitivity to ownership and use costs;
- sensitivity to changes in the level of company cars;
- variation in saturation levels by area and household type;
- employment effects; and
- multi car households (3+)

The new models are applied using a new methodology known as prototypical sampling. This method allows the application of the ownership models to each NTEM zone taking into consideration changes in demographic characteristics for each forecast zone.

With regard to the sample enumeration technique, the method has proved successful in generating artificial samples that are, as far as our target variables suggest, representative of each NTEM zone without diverging too far from the base distribution of household types.

With regard to overall performance, the model has plausible properties with respect to income and price elasticities and ownership forecasts derived from the model compare favourably with actual ownership information extracted from the 1991 census.

The final models have been incorporated in a computer program operating under a WINDOWS environment. A copy of the user manual for this software is appended and included within the program as online help.

## References:

Beckman, R.J., Baggerly, K.A. and McKay, M.D. (1996) "Creating Synthetic Baseline Populations", *Transp. Res. A*, **30**, pp. 415-429.

Daly, A.J. (1999) "How much is enough? Saturation effects using choice models", *Traffic Engineering and Control*

Daly, A.J. (1998) "Prototypical Sample Enumeration as a Basis for Forecasting with Disaggregate Models", PTRC Summer Annual Meeting

Daly, A.J. (1976) "Model Split Models and Aggregation", Local Government OR Unit *Transportation Working Note*

Daly, A.J. and Gunn, H.F. (1985) "Cost-Effective Methods for National-Level Demand Forecasting", IATBR Conference, Noordwijk

Gaudry and Dagenais (1977) "The Dogit Model", Centre de recherche sur les transports, University of Montreal, Canada, Publ. 82, October.

Gunn, H.F. (1984) "Artificial Sample Applications for Spatial Interaction Models", Colloquium Vervoersplanologisch Speurwerk, The Hague

Whelan, G.A. (1999) "A Recalibration of the NRTF Car Ownership Models" Report to the DETR.



## **APPENDIX 1 - DEFINITIONS**

## **Household Income Group Classification (£s pa, 1996)**

Income Group 1 <£5000  
Income Group 2 >£4999 and <£10000  
Income Group 3 >£9999 and <£15000  
Income Group 4 >£14999 and <£20000  
Income Group 5 >£19999 and <£25000  
Income Group 6 >£24999 and <£30000  
Income Group 7 >£29999 and <£35000  
Income Group 8 >£34999 and <£40000  
Income Group 9 >£39999 and <£45000  
Income Group 10 >£44999 and <£50000  
Income Group 11 >£49999 and <£55000  
Income Group 12 >£54999 and <£60000  
Income Group 13 >£59999 and <£65000  
Income Group 14 >£64999 and <£70000  
Income Group 15 >£69999

## **Household Type**

HH1 One adult, not retired  
HH2 One adult, retired  
HH3 One adult, with children  
HH4 Two adults, retired  
HH5 Two adults, no children  
HH6 Two adults, with children  
HH7 Three adults, no children  
HH8 Three adults, with children

## **Area Type (FES)**

Area1 Greater London  
Area2 Metropolitan Districts  
Area3 Districts with density greater than 7.9 persons per hectare  
Area4 Districts with density between 2.22 and 7.9 persons per hectare  
Area5 Districts with density less than 2.22 persons per hectare.

## **Area Type (NTS)**

Area1 London Boroughs  
Area2 Metropolitan Districts  
Area3 Districts with density greater than 10.0 persons per hectare  
Area4 Districts with density between 2.0 and 10.0 persons per hectare  
Area5 Districts with density less than 2.0 persons per hectare.

## **APPENDIX 2 – OWNERSHIP PLOTS**

**APPENDIX 3: A NOTE ON METHODOLOGY FOR PREPARING THE BASE  
INCOME DATA FOR EACH ZONE**

NTEM Zones are specified as agglomerations of Local Authority Wards as constituted at the time of the 1991 Census. Data on household income, certain associated Census variables required for cross-checking it against the principal data-set and selected updates to 1995 are all held by Passenger Transport Networks [PTN] on a Postal Sector basis (that being the more appropriate geography for consumer studies because of the flexibility of the Postcode system). A method of conversion was therefore required.

The first method PTN tried used a nearest-match approach. In the Postal Address File [PAF] supplied by Royal Mail through MapInfo Ltd each of the 1.6 million current (1999) Unit Postcodes is allocated to a current Local Authority Ward. With the assistance of DETR the Ward Codes were adjusted to those for 1991 Wards, although different coding systems and some errors in the files made this a complex task. In a small number of cases of post-1991 changes some arbitrary partitioning was necessary. In PAF each Postcode is located by a 12-character Ordnance Survey Grid Reference [OSGR] and has a count of the number of Residential Delivery Points [RDPs] it contains. The RDPs, as a proxy for population, were used to weight the OSGRs in order to derive a centroid for each Ward. By means of a linking table the Wards were then further accumulated into NTEM Zones and a weighted Zone centroid calculated .

The next step was to run a simple program (in dBASE) in which each Postal Sector was allocated to a Zone by searching for the Zone centroid nearest to the Sector centroid (using Pythagoras' Theorem and the OSGRs). This procedure is equivalent to a map-based search in which each Zone is defined as a Thiessen polygon around its centroid, and each Sector is allocated to the Zone within whose polygon its centroid falls. Finally the data was summed from the Sector level to Zone totals. using the SQL facility in *MapInfo*.

In the initial absence of boundaries or centroids for the Zones this method was felt to be the only feasible one. However, cross-checking against Census data obtained directly on a Ward basis indicated unacceptably large differences. It was thought that these were most likely to be caused by the fact that real and simplified polygonal boundaries do not coincide at all closely, with the result that a number of sectors were misallocated. This was borne out later when it was found that the discrepancies in *ratios* such as people/household and cars/household were much less pronounced than the discrepancies in the *absolute* population figures.

The second method therefore used real boundaries: because the Zones had been specified as lists of Wards it had not been appreciated that Zone boundaries do exist in digitised format. DETR made available a sample file for one Region. Using standard Geographic Information System allocation tools in *Mapinfo* (ie. an SQL enquiry based on "Sector centroid contained within boundary") the Census and income data was summed for each Zone. In this test the population counts were almost identical to those obtained from the Census files, suggesting that mismatching has been largely eliminated. The Income data for each Zone can therefore be incorporated in the models with some confidence. Unfortunately, a glitch in transmission meant that the complete (and very large) file of digitised boundaries supplied by DETR was not ready in time for the first stage of work, but it will be employed in subsequent stages.

The Income data was acquired by ITS in the course of a previous study and relates to 1996. If it proves to be an important element in the present forecasting exercise more up-to-date figures are now available from several sources. Gross Household Income is modelled from a variety of Census variables (of which car-ownership was only a small component, hence minimising any risk of circularity) and from data generated by large-scale consumer market-research. The methodology and results are thought to be robust for the present application.

## **APPENDIX 4 – USER MANUAL**

(This user manual is included in the program and is available on-line)