Skip all front matter: Jump to Page 1

This product is part of the RAND Corporation technical report series. Reports may include research findings on a specific topic that is limited in scope; present discussions of the methodology employed in research; provide literature reviews, survey instruments, modeling exercises, guidelines for practitioners and research professionals, and supporting documentation; or deliver preliminary findings. All RAND reports undergo rigorous peer review to ensure that they meet high standards for research quality and objectivity.

TECHNICAL
REPORT

Sydney Strategic Model Population Synthesiser, 2006 Base

James Fox, Andrew Daly, Bhanu Patruni,
Flavia Tsang

Prepared for the Bureau of Transport Statistics, Transport for NSW

RAND EUROPE

The research described in this report was prepared for the Bureau of Transport Statistics, Transport for NSW.

Published 2012 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
Westbrook Centre, Milton Road, Cambridge CB4 1YG, United Kingdom
RAND URL: http://www.rand.org/
RAND Europe URL: http://www.rand.org/randeurope
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Preface

RAND Europe was commissioned by the Bureau of Transport Statistics (BTS) of Transport for NSW (New South Wales) to update the population synthesiser used in the Sydney Strategic Model (STM) to reflect a 2006 base year.

The STM was designed by Hague Consulting Group in 1997. In Stage 1 of model development (1999-2000), Hague Consulting Group developed mode-destination and frequency models for commuting travel, as well as models of licence ownership and car ownership. In addition a forecasting system was developed incorporating these components. In Stage 2 of model development (2001-02), RAND Europe, incorporating Hague Consulting Group, developed mode and destination and frequency models for the remaining home-based purposes, as well as for non-home-based business travel. During 2003-04, RAND Europe undertook a detailed validation of the performance of the Stage 1 and 2 models, and updated the base year of the Population Synthesiser module to 2001. In 2007 Halcrow undertook Stage 3 of model development, in which they re-estimated the home–work mode-destination models, and at the same time developed models of access mode choice to train for home–work travel.

By 2009, some model parameters dated back to 1999, raising concerns that the model may no longer reflect with sufficient accuracy the current behaviour of residents of Sydney. Furthermore, changes to the zone structure of the model resulted in a trebling of the number of zones as a result of using smaller zone sizes, and because the model study area has been extended to include Newcastle and Wollongong. Therefore, in 2009 RAND Europe was commissioned to re-estimate the STM models.

In this stage of model development (2009-10), the models estimated during Stages 1 to 3 were re-estimated using more recent Household Travel Survey data in order to reflect travel conditions for a new 2006 base year. Furthermore, the scope of the mode-destination models was extended in two ways. First, these models explicitly incorporated the choice between tolled and non-tolled alternatives for car driver travel. Second, access mode-choice models were estimated across a number of travel purposes, and were extended so that the choice of station zone for park-and-ride and kiss-and-ride travel was represented for the first time.

Two reports were produced by RAND Europe during the course of the re-estimation work:

- a mode-destination modelling report (DRR-5270-BTS); and
- a frequency, licence, car ownership modelling report (DRR-5271-BTS).

This report describes work undertaken during 2010-11 to update the Population Synthesiser used in the STM to use new licence and car ownership model parameters, reflecting a 2006 base year, and to use a new process to improve the spatial forecasts of car ownership. The outputs from the population synthesiser are used as inputs to the application system. In future it is envisaged that BTS staff will update the population synthesiser regularly as new data becomes available, and therefore an important role of this report is to explain to BTS staff how the Population Synthesiser works, and to provide a supporting technical manual to facilitate the BTS in running the system and updating the system with new data over time.

Contents

Table of figures

Table of tables

Acknowledgements

The project team acknowledges the contribution of Frank Milthorpe, who has made a substantial contribution to the development of the STM over time. In this project, Frank has provided data for analysis, and helped to produce a number of the spatial plots in this report.

The project team would also like to acknowledge the contribution of the two quality assurance reviewers, Charlene Rohr and Dimitris Potoglou, whose comments have improved the clarity and accessibility of this report.

CHAPTER 1 Introduction

This chapter provides background information for this study. Section 1.1 provides some background to this work and then Section 1.2 summarises the objective of this engagement. Section 1.3 summarises the structure of the four key components of the Sydney Strategic Model, and summarises how the work documented in this report relates to the other three components. The final section of the introduction sets out the structure for the remainder of the report.

1.1 Background

The Bureau of Transport Statistics (BTS) of Transport for NSW operates the Sydney Strategic Travel Model (STM) to inform long term transport planning, policy development and infrastructure assessment in Greater Sydney.

The STM is an advanced tour-based model incorporating extensive market segmentation. The model parameters were initially estimated in the late 1990s and early 2000s. It is implemented using a number of components, using EMME, Excel, ALOGIT and customised FORTRAN programs.

RAND Europe recently re-estimated the model parameters for the STM using more recent data from the Household Travel Survey (see Fox *et al*, 2010, and Tsang *et al*, 2010). Changing to the 2006 zone system for the model has resulted in a trebling of the number of zones as a result of using smaller zone sizes, and because the model study area has been extended to include Newcastle and Wollongong. The complexity of the STM mode destination choice model has increased with the incorporation of train access and car toll nests in the structure. A parallel project has been commissioned to update the ALOGIT implementation of the mode and destination choice models.

An important change that impacts on this work, and the parallel project to update the travel demand models, is that there have been some changes to the segmentation within the model, such as dropping the manufacturing/non-manufacturing worker split from the home–work model, and revisions to the definitions of the income segmentations. A related issue is the move to a 2006 base year, which means that income categories are now defined in 2006 prices, whereas in the previous 2001 base version of the STM incomes were defined in 1996 prices.

1.2 Objectives

The main objective of this work was to update the Population Synthesiser used in the STM to use the new demand model structures and parameters and to reflect a 2006 base year. The outputs from the new 2006 base Population Synthesiser will be used as inputs to the 2006 base application system.

In future, it is envisaged that BTS staff will update the Population Synthesiser regularly as new data becomes available. Therefore, this report was been written with the objectives of both explaining how the Population Synthesiser works and how the system has been updated to a 2006 base, as well as providing sufficient detail to enable BTS staff to run the model and update it over time. A technical manual is provided in Appendix C to help achieve this final objective.

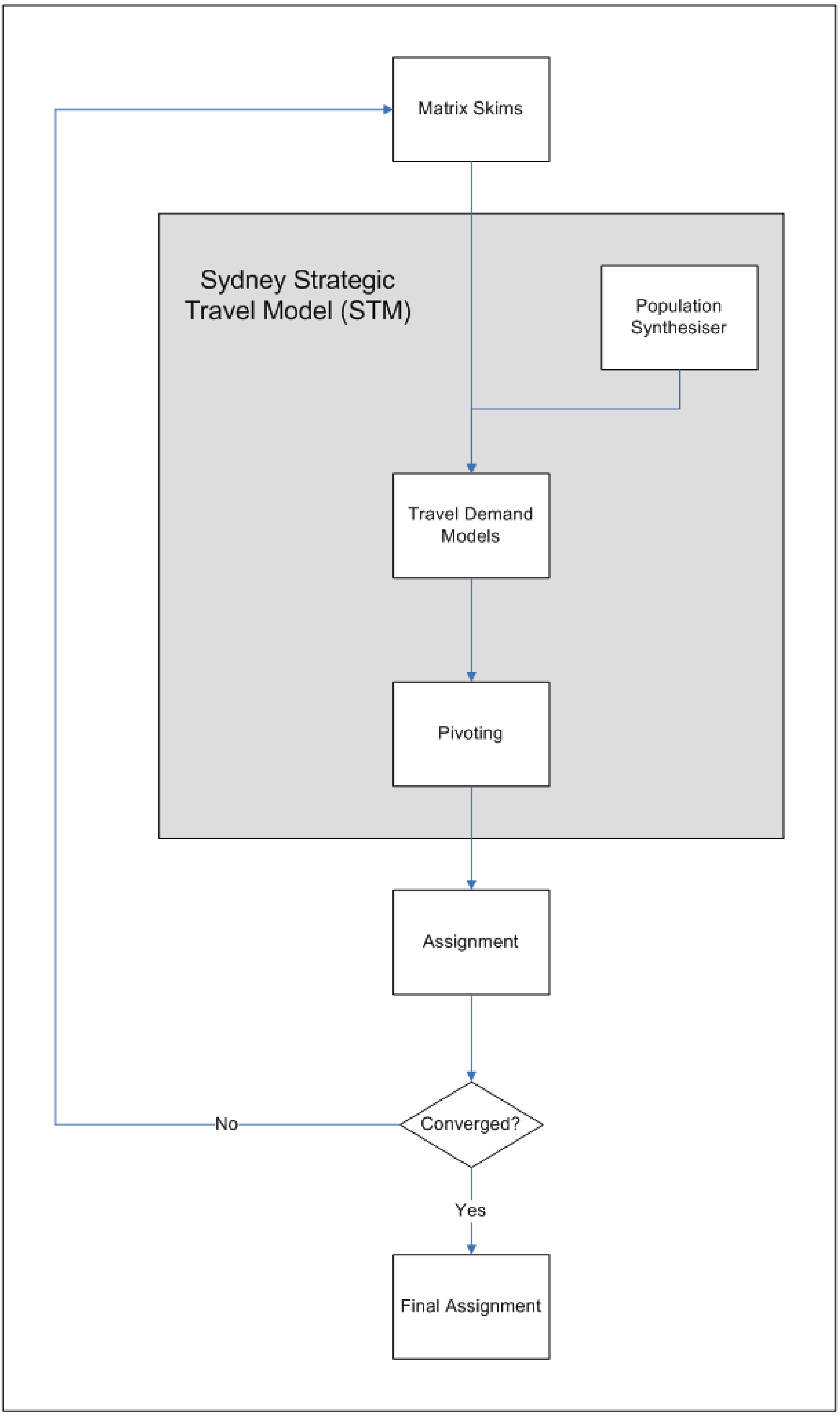
1.3 Structure of the Sydney Strategic Travel Model

The Sydney Strategic Travel Model (STM) system comprises four main components:

1. the Travel Demand Models, which have been updated in a parallel project;
2. a Population Synthesiser, the focus of this report;
3. a pivoting procedure, used to apply predicted changes in travel forecast by the Travel Demand models to base highway and public transport matrices; and
4. network assignments, run separately in EMME for highway and public transport, that have already been updated for the new STM.

The linkage between these four components is illustrated overleaf in Figure 1. It can be seen that the system is run iteratively in order that an acceptable level of convergence is achieved between supply (as represented by the EMME network assignments) and demand (represented by the Travel Demand models).

Figure 1: Structure of the Sydney Strategic Travel Model



1.4 Structure of this report

Chapter 2 describes the Population Synthesiser, and shows how the different components relate to one another. It also describes the updates that have been made to the Population Synthesiser that are specific to each particular component.

Chapter 3 describes updates to the Population Synthesiser that apply to a number or all of the components. Sections discuss the updated data, the treatment of income, the use of area types, and revisions to the definitions of household categories.

Chapter 4 describes an extension to the scope of the Population Synthesiser so that the full spatial variation in car ownership, as measured by the 2006 Census, is captured in the forecasts of base and future car ownership levels.

Chapter 5 describes the model runs and analysis that have been undertaken to test the QUAD component of the model system for the 2006 base year and for seven forecast years (2011, 2016, 2021, 2026, 2031, 2036 and 2041). QUAD is the program that applies the prototypical sampling procedure which expands the base sample the targets variables defined for base and forecast years.

Finally, Chapter 6 presents a summary of the work to update the Population Synthesiser.

Appendix A details changes to the household categories used in the prototypical sampling.

Appendix B defines the four area types used in the Population Synthesiser.

Appendix C is the technical manual for the 2006 base Population Synthesiser.

A separate volume of appendices has been produced, titled 'QUAD GIS validation plots', which presents plots that validate the predictive performance of the 'QUAD' process over spatial areas.

CHAPTER 2 The Population Synthesiser

2.1 Overview

The Population Synthesiser originally comprised five components (HCG and ITS, 2000):

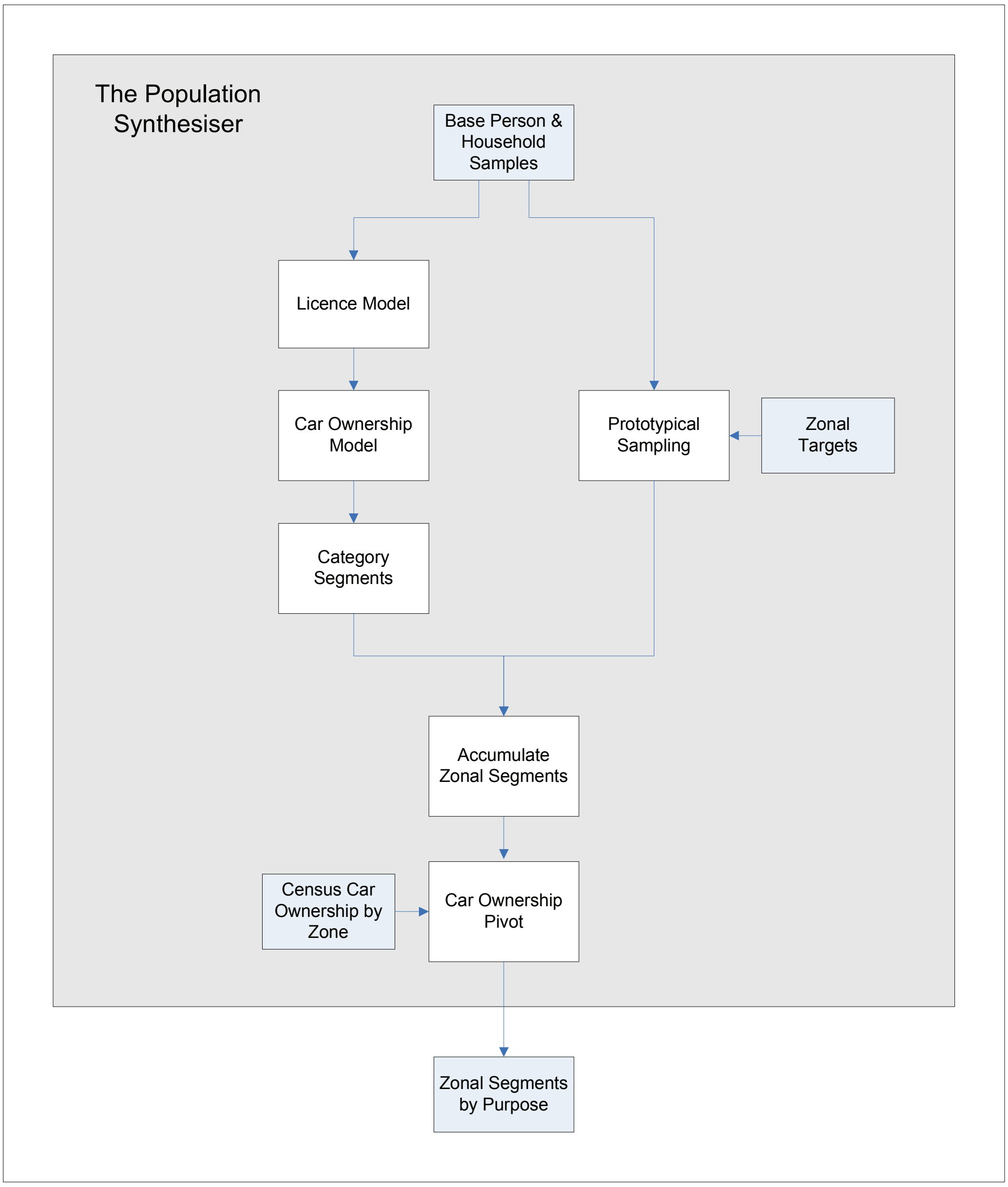
1. licence model;
2. car ownership model;
3. classification into category segments;
4. prototypical sampling procedure; and
5. accumulation of zonal segments.

As part of this work, an additional sixth component has been added to this list:

6. car ownership pivot.

The linkage between these six components is illustrated in Figure 2 overleaf. The first five components are then explained in Sections 2.2 to 2.6. The new car ownership pivot component is described in Chapter 4.

Figure 2: The 2001 Population Synthesiser



The output from the process, Zone Segments by Purpose, defines for each home-based model purpose the population by model zone and segment.

Figure 2 shows the licence model, car ownership model and category segments process as being run in parallel with the prototypical sampling procedure. This was the structure of the 2001 base Population Synthesiser, and reflects the fact that the licence and car ownership model implementations take place at the study area level. However, in the new 2006 base Population Synthesiser, for future year runs it is necessary to run the prototypical sampling first in order to determine a factor used to increase incomes on the base person and household samples. This issue is discussed further in Section 3.3.

2.2 Licence model

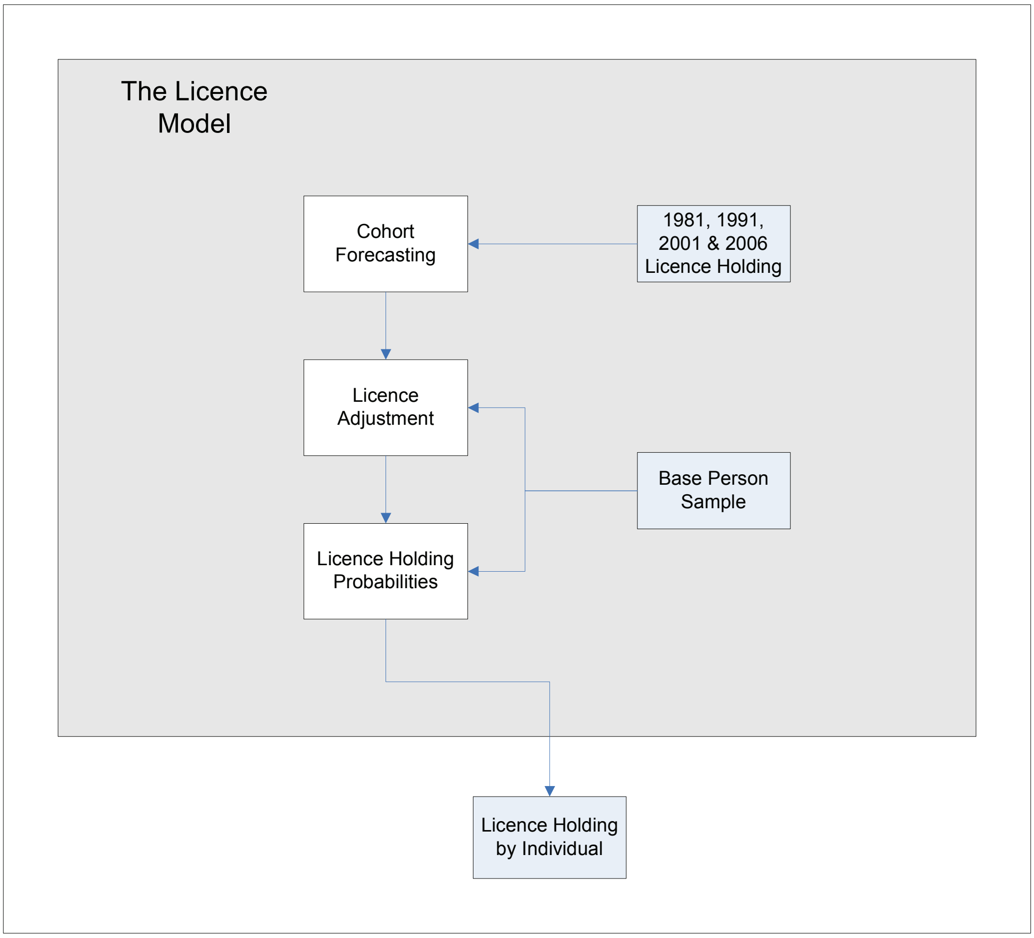
The licence model is comprised of three sub-models:

- cohort forecasting of total licence holding;
- licence adjustment to apply cohort adjustments to disaggregate licence holding models; and
- application of adjusted licence holding models to prototypical sample.

The cohort model is implemented as an Excel spreadsheet. The other two sub-models are coded in the ALOGIT software.

The linkage between these three sub-models is illustrated in Figure 3. The sub-sections following Figure 3 then describe each of the sub-models in turn.

Figure 3: The licence model



2.2.1 Cohort forecasting

The cohort forecasting procedure is undertaken using a spreadsheet system, which was updated as part of the recent re-estimation work (Tsang *et al*, 2010). The spreadsheet was extended to use additional 2006 data and to project forward to 2041 as part of the re-estimation work. Therefore, no further work was required to update the spreadsheet in this project.

The output of the cohort forecasting procedure is predictions of aggregate licence holding by 32 age-gender cohorts for the forecast year.

2.2.2 Licence adjustment

The licence adjustment procedure reads in the forecasts of the cohort procedure, and adjusts the disaggregate licence holding models so that they are consistent with the aggregate cohort forecasts. The licence adjustment procedure is coded in the ALOGIT software.

The licence adjustment procedure has been updated to:

- read in 2006 base person data from the Household Travel Survey to reflect the 2006 base year (documented in Section 3.1);
- read in the parameters from the new licence holding models, and to reflect the changes to the licence holding model specifications described in Tsang *et al* (2010); and
- adjust costs to 2006 prices, and implement the welfare factor adjustment (described in Section 3.3.2).

The licence adjustment procedure developed in the Stage 1 estimation work applied the cohort forecasts by adjusting licence holding for eight age-gender cohorts. The four age bands used were:

- under 25
- 25 to 40
- 41 to 60
- over 60

Taking account of the recent trend for individuals aged under 35 to delay licence acquisition (Raimond and Milthorpe, 2010), and to take account of higher licence holding up to the age of 70 due to cohort effects (Tsang *et al*, 2010), the cohorts were revised and a new cohort added, giving five in total:

- under 25
- 25 to 34
- 35 to 54
- 55 to 69
- 70 plus

The output from the licence adjustment procedure is a set of adjusted model parameters for the licence holding models.

2.2.3 Licence holding models

The licence holding models read in the licence holding adjustment parameters, which have been adjusted to take account of the cohort forecasts for the five age bands and by gender, and apply them to the base person sample to predict the licence holding of each adult in the sample. The licence holding models are coded in the ALOGIT software.

It is noted that there are actually two licence holding models, one for the head of the household and their partner, if they exist, and one for any other adults in the household. Both models have been updated for this and the licence adjustment sub-tasks.

The two licence holding models have been updated to:

- read in 2006 base person data from the Household Travel Survey to reflect the 2006 base year (documented in Section 3.1);
- read in the parameters from the new licence holding models, and to reflect the changes to the licence holding model specifications described in Tsang *et al* (2010); and
- adjust costs to 2006 prices, and implement the welfare factor adjustment (see Section 3.2.3).

The output from the licence holding models is a file outputting the licence holding probability of each adult in the base sample. For the head of the household and their partner, if they exist, probabilities are output for each of the four licence holding combinations[1]. For any other adults in the household, licence holding probabilities are output for the same combinations of head and partner licence holding, because the other adults' model contains terms for the licence holding of the head and their partner. Licence holding probabilities are output for all the adults in the household because licence holding terms are used in the company and total car ownership models which are applied conditional on the outputs of the licence holding model.

2.3 Car ownership model

The car ownership model comprises two sub-models, both of which operate at the household level:

1. a company car ownership model; and
2. a total car ownership model, which is applied conditional on the predictions of the company car ownership model.

Both sub-models are coded in the ALOGIT software, and have been updated as part of this work. The linkage between these two sub-models is illustrated in Figure 4. The updates that have been made to the two models are described in Sections 2.3.1 and 2.3.2.

A new component has been added to the car ownership model during this work to pivot off the car ownership levels observed in the 2006 Census. This component is the last

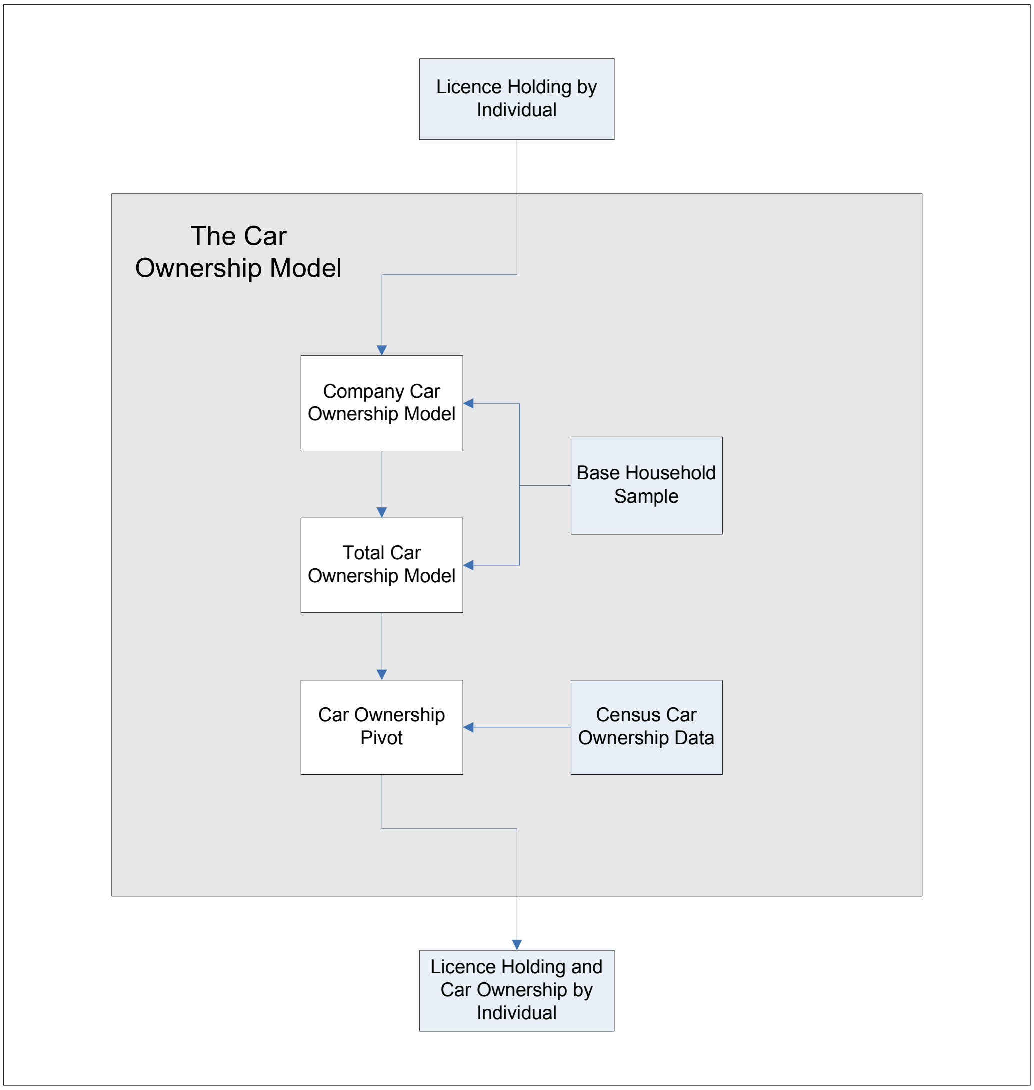
---

[1] Head owns a licence, partner owns a licence, both own licences, neither own licences.

component of the Population Synthesiser to be run, and is documented separately in Chapter 4.

Section 2.3.3 describes the updates that have been made to the car availability adjustment procedure. This procedure adjusts the availability of cars at the person level to take account of changes in accessibility. This procedure is not shown in Figure 4 as it implemented within the Travel Demand models, rather than as part of the Population Synthesiser.

Figure 4: The car ownership model



2.3.1 Company car ownership model

The company car ownership model reads in the licence holding probabilities for each individual in the base sample, and calculates company car ownership probabilities associated with each licence holding state for households with at least one worker. The options are no company car, one company car and two-plus company cars. For households with no workers the probability of owning a company car is set to be exactly zero.

It is noted that the company car ownership model is a household level model. Therefore, the company car ownership model calculates car ownership at a household level taking information from the different persons that comprise each household. These calculations are made separately for each possible combination of head and partner licence holding, and take account of the licence holding probabilities of any other adults in the household, as the probability of company car ownership depends on licence holding.

The company car ownership model is coded in the ALOGIT software.

The company car ownership model has been updated to:

- read in 2006 base household data from the Household Travel Survey (see Section 3.1);
- read in the parameters from the new company car ownership model, and to reflect the changes to the model specification described in Tsang *et al* (2010); and
- adjust costs to 2006 prices, and implement the welfare factor adjustment (see Section 3.2.3).

The output from the company car ownership model is a file detailing, for each household, the probability of each of the three company car ownership outcomes for each possible combination of head and partner licence holding.

2.3.2 Total car ownership model

The total car ownership model reads in the licence holding probabilities of each individual in the base sample, and the associated company car ownership probabilities, and calculates the car ownership probabilities associated with each possible licence holding and car ownership state. The options are no car, one car, two cars and three-plus cars.

The total car ownership model is coded in the ALOGIT software.

The total car ownership model has been updated to:

- read in 2006 base household data from the Household Travel Survey;
- read in the parameters from the new total car ownership model, and to reflect the changes to the model specification described in Tsang *et al* (2010); and
- adjust costs to 2006 prices, and implement the welfare factor adjustment (see Section 3.2.3).

The output from the total car ownership model is a file detailing, for each household, the probability of each of the four total car ownership outcomes for each possible combination of head and partner licence holding and company car ownership outcome. As per the company car ownership model, the licence holding of any other adults in the household is taken into account when calculating the total car ownership outcomes.

2.3.3 Car availability adjustment

The car availability adjustment adjusts total car ownership and thus the availability of cars at the person level to take account of changes in home–work accessibility in the forecast year relative to the base year. In the travel demand models, car availability segments are defined using information on car ownership together with individual and household licence holding, and these car availability segments govern the availability and attractiveness of the car alternatives in the travel demand models.

The car availability adjustment process is iterative, and is applied as part of the standard iteration between the travel demand models and the EMME assignments. Because of this link to the assignment process, it is implemented as part of the travel demand models, and not as part of the Population Synthesiser (and hence it does not appear in Figure 4). However, the updates to the car availability procedure were commissioned as an optional component of the Population Synthesiser work, and for that reason the updated process is documented in this report.

The procedure employs a simple pivot-point model[2] to predict changes in the probability of car availability states as a function of the change in home–work accessibility relative to the base year. The parameters used in the pivot-point model in the 2001 base version of the STM were calibrated during the Stage 2 work in 2001-02. More information on the car availability adjustment procedure, including specification of the pivot-point model, is provided in Chapter 3 of HCG and ITS (2002).

The new pivot point parameters for the 2006 base version of the STM have been calibrated by making a series of runs of the base home–work Travel Demand model with 10% changes applied to the home–work accessibility for particular car availability segments. By assessing the predicted changes in the distribution of persons across the extended car availability segments (*aext2*), the new pivot-point parameters can be calculated.

Table 1 defines the eight home–work car availability segments a, and shows how they map to the 16 extended car availability segments *aext2*. The 16 extended car availability segments *aext2* are then defined in Table 2. Competition for car means that the number of household licences is greater than the number of cars in the household. Free car use means that the number of household licences is less than or equal to the number of cars in the household.

Table 1: Home–work car availability segments

a	Definition	*aext2*
1	No car in household	1, 2, 3, 4
2	No personal licence, but car in household	5, 10, 11, 16
3	Competition for car, no company car	6, 12
4	Free car use, one household licence, no company car	7
5	Free car use, several household licences, no comp. car	13
6	Competition for car, 1+ company car	8, 14
7	Free car use, one household licence, 1+ comp. car	9
8	Free car use, several household licences, 1+ comp. car	15

[2] A 'pivot-point' model predicts changes from a known base situation on the basis of differences in utility between the forecast and base cases.

Table 2: Extended car availability segmentation *aext2*

aext2	Definition
1	No cars, no licence in household
2	No cars, personal licence, one licence in household
3	No cars, no personal licence, at least one licence in household
4	No cars, personal licence, two plus licences in household
5	No personal licence, one non-company car
6	Competition for car, one non-company car
7	Free car use, one household licence, one plus non-company car[3]
8	Competition for car, one company car
9	Free car use, one household licence, one plus company car[3]
10	No personal licence, one company car
11	No personal licence, two plus non-company cars
12	Competition for cars, two plus non-company cars
13	Free car use, several household licences, no company car
14	Competition for cars, one plus c.car, two plus cars in total
15	Free car use, several household licences, one plus company car
16	No personal licence, one plus c.car, two plus cars in total

As a consequence of home–work accessibility changes, and hence changes in total car ownership, the car availability adjustment procedure forecasts changes within the following groups of *aext2* codes:

aext2 = 1 no licences in household, such households can never acquire cars

aext2 = 3, 5, 11 no licence, but at least one household licence, no company cars

aext2 = 10, 16 no licence, at least one household licence, at least one company car

aext2 = 2, 7 individual has only licence in household but no company cars

aext2 = 9 individual has only licence in household and a company car

aext2 = 4, 6, 12, 13 individual and at least one other person in household have licences, no company cars

aext2 = 8, 14, 15 individual and at least one other person in household have licences, at least one company car

It can be seen that there are five groupings for which shifts between *aext2* segments are possible in response to changes in accessibility (for *aext2* codes 1 and 9 no shifts are possible). For each of these five groups, a base *aext2* segment is defined and pivot-point parameters are defined to specify the shifts from *aext2* segments relative to the base segment that occur as a result of the accessibility changes.

The accessibility changes are applied to the home–work car availability segments a. Changes to car availability segments 2 and 7 have no impact on personal level car availability, and so to calibrate the pivot-point parameters six runs of the home–work TravDem have been made, applying 10% changes in logsums to segments 1, 3, 4, 5, 6 and 8 respectively.

[3] It is assumed that for these segments multiple car dummies are not applicable, as there is only a single licence holder to drive the cars.

The new model parameters are summarised in Table 3, where the new 2006 values are compared to those used in the 2001 base version of the STM. The following information is presented in Table 3:

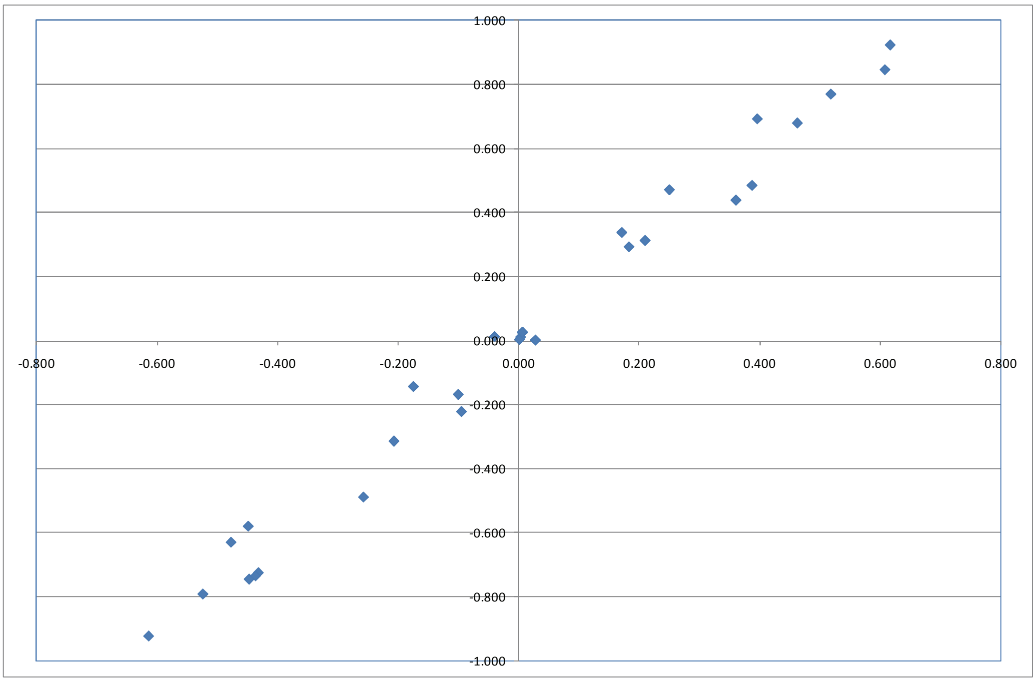
- the base *aext2* segment (the *aext2* segments are defined in full in Table 16);
- the pivot point coefficient (e.g. c05_1 defines the shifts for *aext2* segment 5 relative to the base category 3 resulting in changes to the logsum applied to home–work car availability segment 1);
- the value of the pivot point coefficient obtained in the new 2006 base version of the STM;
- the value of the pivot point coefficient obtained in the previous 2001 base version of the STM; and
- the absolute value of the ratio of the 2006 and 2001 base coefficients.

Table 3: Car availability adjustment model parameters

Base aext2	Coeff	STM 2006	STM 2001	abs(06/01)
3	c05_1	-0.47808	-0.62994	0.759
	c05_3	-0.44948	-0.57988	0.775
	c05_4	0.18295	0.29373	0.623
	c05_5	-0.09532	-0.22162	0.430
	c11_1	0.38720	0.48543	0.798
	c11_3	0.36046	0.43932	0.820
	c11_4	-0.10072	-0.16781	0.600
	c11_5	0.17104	0.33850	0.505
10	c16_6	-0.25807	-0.48879	0.528
	c16_8	0.24995	0.47186	0.530
2	c07_1	-0.52490	-0.79180	0.663
	c07_3	0.00283	0.01134	0.249
	c07_4	0.51813	0.77095	0.672
	c07_5	0.00078	0.00355	0.220
4	c06_1	-0.44801	-0.74531	0.601
	c06_3	0.60794	0.84708	0.718
	c06_4	0.00615	0.02676	0.230
	c06_5	-0.17520	-0.14313	1.224
	c12_1	-0.43259	-0.72501	0.597
	c12_3	0.46258	0.68054	0.680
	c12_4	0.00585	0.02576	0.227
	c12_5	-0.04036	0.01262	3.198
	c13_1	-0.43737	-0.73512	0.595
	c13_3	0.02751	0.00199	13.808
	c13_4	0.00588	0.02595	0.227
	c13_5	0.39596	0.69331	0.571
8	c14_6	-0.20740	-0.31383	0.661
	c14_8	0.20969	0.31333	0.669
	c15_6	-0.61478	-0.92319	0.666
	c15_8	0.61686	0.92451	0.667

It is difficult to directly interpret the pivot model parameters, as their magnitudes depend on the scale of the logsum accessibility measures, which varies between the 2001 base and 2006 base versions of the STM. Therefore, a scatter plot was generated to compare the old and new models parameters, presented in Figure 5. The correlation between the old and new parameters is 0.993.

Figure 5: Car availability adjustment model parameters scatter plot



It is clear from the scatter plot, and the correlation measure, that the old and new model parameters are closely related.

2.4 Classification into category segments

This process takes the household-person files output by the licence and car ownership models, and classifies records into the segments used in the mode-destination and frequency models, and household categories used for the prototypical sampling procedure.

For a given record, the weight is given by the product of:

- the probability of the licence ownership state;
- the probability of the company car ownership state;
- the probability of the total car ownership state; and
- the person weight.

The person weight is the expansion factor from the Household Travel Survey (HTS) that expands the base sample to reflect the 2006 base population. Multiplying the person weight by the probability of a particular licence holding and car ownership state gives the predicted number of persons in each possible state. Considering each possible state allows the number of persons in each car availability segment to be predicted.

The process has been updated to reflect the changes to the mode-destination and frequency segments that are documented in Chapter 3 of Fox *et al* (2011). Only a few changes were made to the segments, and therefore only relatively minor changes were necessary to update the process that categorises individuals into household categories and segments.

One important change to the segments resulted from the car ownership pivot process. Specifically, a new extended car availability segmentation, *aext3*, was defined to allow the car ownership pivot process to operate. The *aext3* segmentation is documented in Section 4.4.

The output from the category segment process is a person file that classifies the record into each of the mode-destination and frequency segments, and details the weight associated with that record.

2.5 Prototypical sampling

The prototypical sampling procedure expands the base sample to best match population targets defined for each zone in the study area. The expansion is undertaken using a quadratic minimisation that is applied for separately for each of the 2,690 zones in the study area:

For each zone, $\min_{\varphi} F(\varphi)$, subject to $\varphi_i \geq 0$ for all categories i, where

$$F(\varphi) = \sum_t w_t^a \left(y_t - \sum_c \varphi_c^a x_{tc}^a\right)^2 + \sum_c \left(\varphi_c^a - f_c^a\right)^2 \qquad (2.1)$$

where φ_c^a gives the frequency in the population of household category c

a is the area type (discussed in Section 3.4)

f_c^a gives the base frequency in the same units of household category c

w_t^a gives the weight attached to meeting target t

y_t^a is the target variable for the zone (the zonal target)

x_{tc}^a is the average quantity of target variable for a household of category c

It can be seen that the QUAD objective function comprises two elements, the first assessing the fit to the zonal targets, and the second the deviation from the base distribution.

It is the φ variables that are optimised. Note that all the weights for divergences from base frequencies, i.e. the term $\Sigma_c (\varphi_c^a - f_c^a)^2$ in Equation 2.1, are assumed to be equal and these are then given the arbitrary value 1, i.e. the weights w for divergences from the targets are defined relative to this arbitrary scale.

An important point should be emphasised at this stage. The second term in $F(\varphi)$ seeks to minimise the difference between the zonal expansion and the base distribution over household categories. If there are fundamental differences between the base distribution and the distribution implied by the target variables then the zonal targets will never be achieved exactly. Rather the objective function achieves a balance between meeting the zonal targets and retaining the characteristics of the base sample.

An important change to the prototypical sampling procedure in the 2006 Population Synthesiser is that the procedure is now run separately by four area types. These area types are defined in Section 3.4.

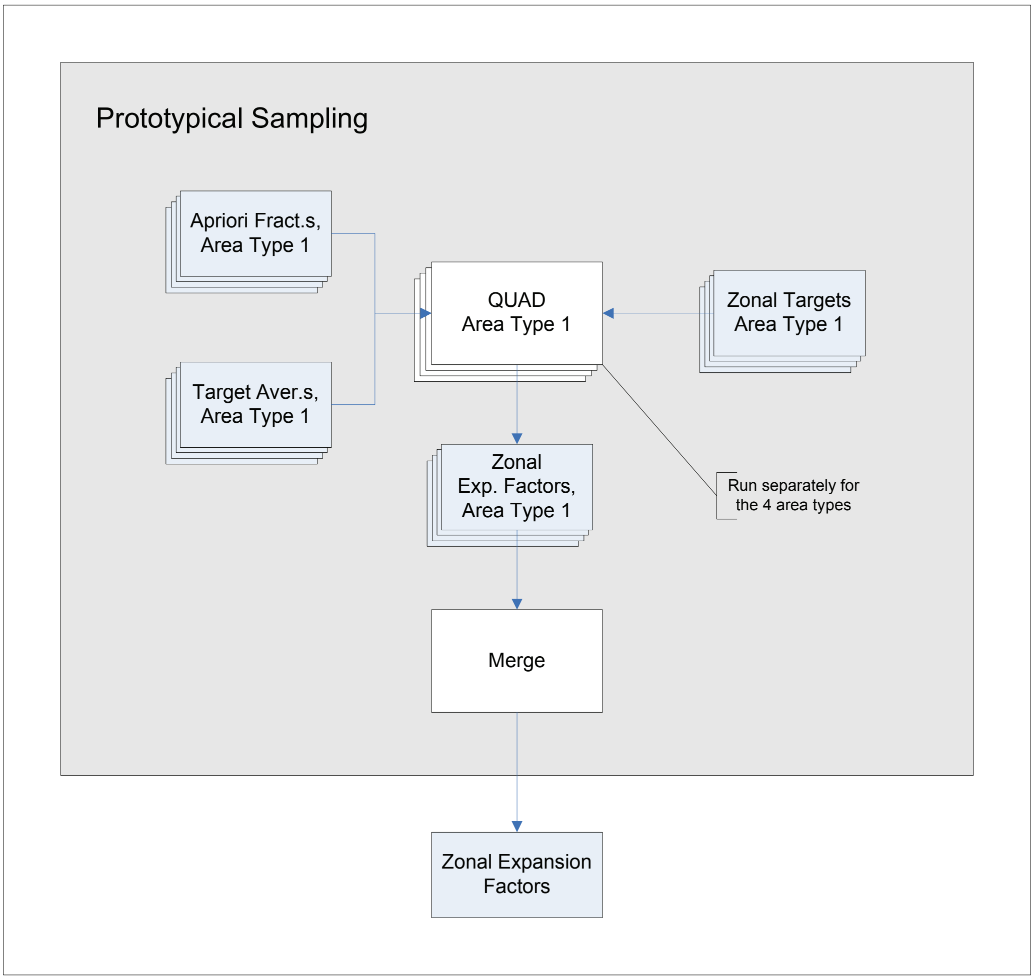
The base sample is drawn from the HTS, and contains both person and household information. It is documented in Section 3.1. The base sample is processed to create the following sets of inputs:

- *A-priori* fractions by area type, which define the base frequencies (number of households) by household category c for a given area type a. They are the terms f_c^a in Equation 2.1.
- Target averages by area type, which define the average value of each target variable t for each household category c for a given area type a. They are the terms x_{tc}^a in Equation 2.1.

The target information for the 2006 base year has been assembled by the BTS from 2006 Census data, and is documented in Section 3.2. The BTS has also made projections of the target data through to 2036, which are documented in Section 3.2. Section 3.2 also provides a definition of the 21 targets used in the 2006 base Population Synthesiser.

Figure 6 illustrates how the different components of the prototypical sampling procedure are linked.

Figure 6: The prototypical sampling procedure



The prototypical sampling process has also been updated to reflect RAND Europe's latest thinking on how to treat income growth in forecasting. As a result, future income growth has been decomposed into a 'welfare effect' applied to all base incomes, and re-distribution between income bands. The treatment of income is discussed in more detail in Section 3.2.3.

Another important change to the input files to the prototypical sampling procedure relative to the previous 2001 base version of the Population Synthesiser is that the process is now run separately for four area types, in order to reflect differences in household characteristics across the study area. The four area types are:

- inner Sydney ring
- middle Sydney ring
- outer Sydney ring
- regional, comprising Newcastle and Wollongong

Section 3.4 presents an analysis demonstrating how the household characteristics vary across the four study areas, and includes maps illustrating the area types.

The procedure operates using the QUAD software, which is a FORTRAN application that was delivered by Hague Consulting Group together with the Stage 1 model in 2000. QUAD uses the Microsoft compiler version of FORTRAN.

The input files to the QUAD procedure have been updated to reflect the new base sample. This required updates to the SPSS syntax used to create the files, which are documented in the technical manual (Appendix C). The SPSS syntax have been commented and documented so that the BTS will be able to update the base sample as new data becomes available.

A new 'Merge' process has been added to the prototypical sampling procedure to merge the outputs from the four QUAD runs by area type into a single output file of zonal expansion factors. The Merge process has been coded in the ALOGIT software.

The output from the prototypical sampling process is then factored for each zone in the study area that expand the household sample to best match the zonal target totals.

2.6 Accumulate zonal segments

The final stage in the population model process is to accumulate the forecast population by segment. This accumulation process happens separately by each of the seven home-based model purposes, because each purpose has different segmentations, and the procedure that implements the process is termed 'ACCUM'.

It is noted that the two non-home-based model purposes, work-based business and non-home-based business detours are implemented as a function of the predictions of the home–work and home–business models. As a result, ACCUM does not provide outputs for the two non-home-based purposes currently represented in the STM.

The procedure operates using the 'ACCUM' software, which is a FORTRAN application that was delivered by Hague Consulting Group together with the Stage 1 model in 2000, and subsequently updated by RAND Europe in 2004 when the Population Synthesiser was updated to use a 2001 base. The version of ACCUM delivered in 2004 used the Salford FORTRAN compiler. The version delivered for the new 2006 base Population Synthesiser uses the Microsoft compiler, which achieves consistency with the compiler used for the QUAD program, and furthermore is the compiler used for the ALOGIT software used for a number of the other Population Synthesiser components.

As well as changing the compiler, the ACCUM program has been modified to reflect the changes to the mode-destination and frequency segments. All but one of the segment changes are detailed in Chapter 3 of the final report for the project, run in parallel with this project, to implement the new mode-destination and frequency models (Fox *et al,* 2011). One other segmentation has been changed in ACCUM, namely the extended car availability segmentation *aext3*. The changes to this segmentation are discussed in Section 4.4.

The output from the ACCUM program is files for each of the seven home-based model purposes that detail the base or future year population by zone and segment.

CHAPTER 3 Updates to the synthesiser

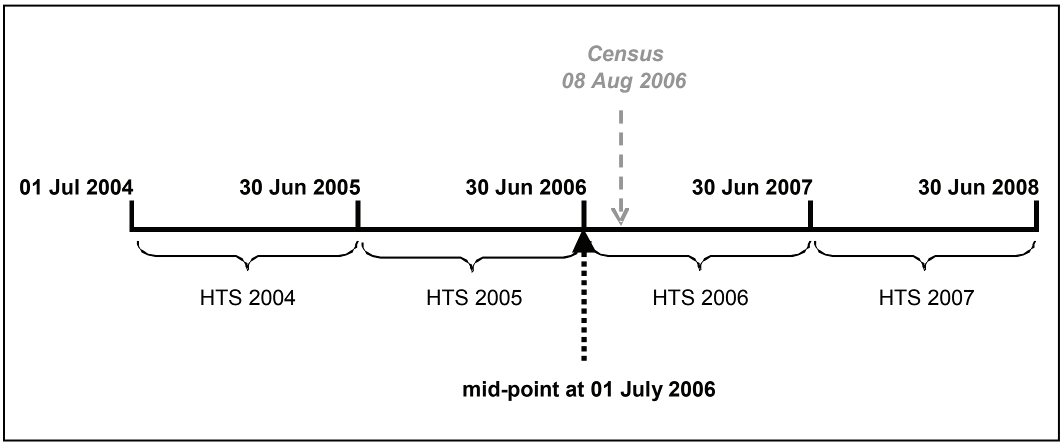
This chapter summarises updates to the model system that have been made during this work that apply to some or all of the components of the Population Synthesiser. It starts by summarising the input data to the Population Synthesiser, with the Household Travel Survey data described in Section 3.1, and the Census target data described in Section 3.2. Section 3.2 also describes revisions to the target definitions used in the prototypical sampling process. Section 3.2.3 describes how incomes, and in particular their growth over time, are treated in the Population Synthesiser. Section 3.4 describes the four area types that have been used in the 2006 base Population Synthesiser to better account for differences in the characteristics of the population across different area types. Finally, Section 3.5 describes revisions to the household categories across which the population is reweighted to match zonal targets.

3.1 Household Travel Survey data

Household Travel Survey (HTS) data is used to form the base person and household sample for the Population Synthesiser. The base sample forms the basis for the zonal expansion undertaken by QUAD. It is also used as input to the licence and car ownership models, which predict licence ownership and car ownership as a function of the person and household information contained on the base sample. Finally, it is used to classify persons into the model segments used in the category segment process.

To define a base sample representative of the 2006 base year, four waves of HTS data has been collected between July 2004 and June 2008. As illustrated in Figure 7, this definition gives a sample of sufficient size that evenly straddles the 2006 base year, and has a mid-point that is close to the data of the 2006 Census data that is used to define the base year zonal targets.

Figure 7: Waves of HTS data used to form base sample



The base sample extracted by the BTS includes households interviewed at weekends as well as on weekdays. Including households interviewed at weekends uses the maximum available data and is correct because only person and household data is used as input to the Population Synthesiser, i.e. trip data is not used as input.

Only 'full response' households are included in the base sample. Full response households are households where all individuals have completed a person form, and therefore there is person information available for each individual in the household.

A further condition that has been applied for the base sample is that all adults in the household must have answered the personal income question. Note that negative or zero income answers are allowed, but missing responses are not. This condition reduced the number of available person records by 12.9%, from 29,555 to 25,745.

A total of 25,745 persons form the base sample, which is drawn from 9,915 households. This represents some reduction in sample size relative to the 2001 base version of the Population Synthesiser, which used five waves of data (collected between July 1997 and June 2002), not four. The change in sample size is shown in Table 4.

Table 4: Change in prototypical sample size

	2001 Base Sample	2006 Base Sample	Percent Change
Persons	32,895	25,745	-21.7 %
Households	12,222	9,915	-18.9 %

The mean household size has reduced from 2.69 in the 2001 base sample to 2.60 in the 2006 base sample.

3.2 Census target data

3.2.1 Target definitions

The 2001 base version of the Population Synthesiser used a total of 19 targets, comprising:

- eight age-gender targets;
- five household type targets;

- two worker type targets; and
- four personal income band targets.

The zonal targets have been revised in this work, both to draw on the findings from the validation work undertaken for the 2001 base version of the Population Synthesiser (Fox *et al*, 2004), and to take account of changes in the definition of income bands.

These are the specific changes to the targets:

- The worker targets for non-manufacturing and manufacturing workers, which have become less relevant over time following the decline in the proportion of the workforce employed in manufacturing, have been replaced with targets for full-time and all other workers (the major component of which is part-time workers).
- A new target to reflect the number of students has been added with the objective of achieving a better fit to the spatial distribution of students in tertiary education across the study area.
- The number of income bands has been increased from four to five, a control target for children has been added (named band 0) to improve the fit for the five income targets for adults, and the definitions of the income bands have been revised.

Table 5 summarises the target definitions in the previous 2001 base and the new 2006 base versions of the population synthesiser. Cells where changes have occurred in the 2006 base version of the Population Synthesiser are highlighted.

Table 5: Comparison of prototypical sampling targets, 2001 base versus 2006 base

Target	Group	2001 Base	2006 Base
1	Age-Gender	Males aged 0–19	Males aged 0–19
2		Males aged 20–39	Males aged 20–39
3		Males aged 40–59	Males aged 40–59
4		Males aged 60+	Males aged 60+
5		Females aged 0–19	Females aged 0–19
6		Females aged 20–39	Females aged 20–39
7		Females aged 40–59	Females aged 40–59
8		Females aged 60+	Females aged 60+
9	Household Types	Couples with children	Couples with children
10		Couples only	Couples only
11		Single parent	Single parent
12		Single person	Single person
13		Other types	Other types
14	Workers	Non manufacturing workers	Full-time workers
15		Manufacturing workers	Part-time workers
16	Income Bands	Band 1 $ 0–6,239	Band 0 children (aged <15)
17		Band 2 $ 6,240–20,799	Band 1 $ 0–20,799
18		Band 3 $ 20,800–41,599	Band 2 $ 20,800–31,199
19		Band 4 $ ≥ 41,600	Band 3 $ 31,200–41,599
20			Band 4 $ 41,600–67,599
21			Band 5 $ ≥ 67,600
22	Students		Tertiary Education Students

It is noted that, in the model code documented in Appendix B, the students target is actually the 16th target, and the income targets are the 17th to 22nd targets. However, the students target is presented last in Table 5 in order to facilitate comparison between the two Population Synthesiser versions.

3.2.2 Base target data

Target data has been assembled by the BTS for the new 2006 base year from Census data.

For zones with fewer than 10 households, a procedure was followed to transfer households from adjacent zones, in order that the target variable distributions were based on a sufficiently reliable sample size. In this process, 'donor' households were transferred from zones with more than 10 households in order to increase the number of households in the recipient zone up to a total of 10. Persons are transferred at the same time as households, and this is done in a way that leaves the mean household size in the transfer zone unaltered. The mean household size in the recipient zone can change, however. Donor and recipient households are always located within the same SLA, which ensures that the household control total for the SLA remains unaltered.

Table 6 illustrates the process of transferring households with a simple two zone example, which shows how the process preserves the household and population totals before and after transferring households, and preserves the household size in the donor zone but not the recipient zone.

Table 6: Transferring households example

Zone	Original			After Transferring		
	HH	Persons	HH size	HH	Persons	HH size
1 (donor)	250	500	2.00	250 – 5 = 245	500 – 5*2 = 490	2.00
2 (recipient)	5	15	3.00	5 + 5 = 10	15+5*2 = 25	2.50
Total	255	515		255	515	

The household transfers to ensure a minimum of 10 households per zone are permanent adjustments. Once these adjustments have been made, the second step is to create the target distributions, which define the *proportions* of persons in the age-sex, worker type and income categories, and the *proportions* of households in the household type categories. For zones with more than 100 households, these distributions are taken directly from the Census information available for that zone. For zones with fewer than 100 households, these distributions borrow information from neighbouring zones so that the distributions are based on a minimum of 100 households. The transfer process is temporary for the purpose of determining the target distribution, rather than the permanent transfer process used for the zones with fewer than 100 households. Note also that a donor zone can contribute its distribution over the target variables to multiple zones with less than 100 households. Once the target distributions have been determined for each zone, they are applied to the number of persons and households in that zone to determine the target totals for that zone. A total of 548 zones were affected by this procedure, 20.4% of the 2,690 zones. However, as these zones contain low numbers of households they contain just 0.8% of the total households in the study area.

As noted in Section 3.2.1, a new tertiary education student target was added for the 2006 base version of the Population Synthesiser. An issue that arose when assembling this data is that a substantial fraction of persons who are reported as students do not specify the type of education facility that they attend, and therefore cannot be classified as either tertiary students or other student types. To overcome this issue, the study wide fraction of students who stated that they attend a tertiary education facility as a proportion of all students who stated an education facility was calculated (16.47%). For each zone, this fraction was then applied to the number of students who did not state their type of education who could have been tertiary students, i.e.:

[Tertiary students] = [Stated as tertiary] + 0.1647 * [No education facility stated]

Table 7 summarises for each of the 21 target variables:

- the mean value of the target variable across the 2690 model zones;
- the sum of the target variable across the 2690 model zones; and
- the percentage of the target variable within the target group.

For the five income targets, the population totals are for persons aged 15 and above only. The totals of the targets across each of the groups are also presented, except for students, as the target total is the target number of students.

Table 7: 2006 Census target data, descriptive statistics[4]

Target	Mean	Total	Percent	
Males, 0–19	253.4	681,554	13.3%	
Males, 20–39	286.3	770,233	15.0%	
Males, 40–59	252.9	680,354	13.3%	
Males, 60+	149.2	401,471	7.8%	
Females, 0–19	241.7	650,277	12.7%	
Females. 20–39	288.4	775,770	15.1%	
Females, 40–59	258.5	695,272	13.5%	Total pop
Females, 60+	177.7	477,953	9.3%	5,132,884
Couples with children	247.6	666,062	34.6%	
Couples only	179.3	482,361	25.0%	
Single parent	82.4	221,754	11.5%	
Single person	167.6	450,884	23.4%	Total hhs
Other types	39.2	105,477	5.5%	1,926,539
Full time workers	688.4	1,851,708	73.3%	Total wkrs
Other workers	250.6	674,137	26.7%	2,525,845
$ 0–20,799	670.1	1,802,505	43.5%	
$ 20,800–31,199	208.1	559,715	13.5%	
$ 31,200–41,599	170.7	459,232	11.1%	
$ 41,600–67,599	269.6	725,306	17.5%	Total adults
$ ≥ 67,600	221.0	594,499	14.4%	4,141,257
Tertiary students	153.5	413,005	8.0%	

[4] Note that the income control target for children (income0) is not presented in Table 7, as the income target distribution has been calculated for adults only.

3.2.3 Future year target data

The BTS has provided a set of future year targets, using the 2006 Census data as the basis.

Table 8 overleaf summarises the future year target data. Within Table 8, four sub-tables are presented:

- the first gives the study-area totals for each of the 21 target totals;
- the second gives the totals by the five target groups, noting that the five income targets sum to the total number of adults, and the tertiary students target forms a group on its own;
- the third gives the percentage growth in the target relative to the 2006 base year, with red highlighting used to indicate where the growth is lower than for that target group as a whole, and green highlighting used to indicate where the growth is higher than for that target group as a whole; and
- the fourth table gives the percentage growth in each target group relative to the 2006 base year.

Review of Table 8 shows that the population of the study area is predicted to grow by around 6% every five years. The targets demonstrate an ageing population, with noticeably higher than average growth predicted for men and women over 60 years of age. The number of females aged 60+ is predicted to almost double by 2041, and the number of males aged 60+ is predicted to more than double.

The total number of households is predicted to grow more rapidly than the total population, and it can be seen that this is due to higher than average growth in couple only, single parent and single person households, with single person households growing most rapidly.

In total, worker growth falls below population growth, as the percentage of the population of working age is predicted to grow less rapidly than population as a whole. This follows as a consequence of the ageing of the population. Growth for the other worker type is much more rapid than growth in the number of full-time workers as a consequence of a prediction of more part-time working.

For the income targets, predicted real terms income growth of 1% per annum results in below average growth in the number of adults in the first three income bands, and above average growth in the number of adults in the top two income bands. Substantial growth is predicted in the numbers of adults falling in the top income band. These changes are illustrated in Figure 8, which is presented after Table 8.

The tertiary students target shows growth slightly below growth in the population as a whole, a result consistent with the ageing of the population.

Table 8: Future year target totals (2011, 2016, 2021, 2026, 2031, 2036, 2041)[5]

Totals by target

Target	2006	2011	2016	2021	2026	2031	2036	2041
Males, 0-19	681,541	709,054	736,487	770,393	806,181	837,234	866,164	905,868
Males, 20-39	770,219	804,539	839,605	876,119	904,572	934,064	963,656	1,006,319
Males, 40-59	680,376	703,455	732,100	749,305	773,990	807,067	840,498	878,723
Males, 60+	401,507	469,577	535,578	608,161	680,705	747,017	811,595	848,656
Females, 0-19	650,263	675,011	700,712	731,795	765,599	794,851	822,133	859,360
Females. 20-39	775,729	811,916	844,978	879,720	905,635	933,955	962,514	1,004,027
Females, 40-59	695,274	722,787	754,037	773,950	802,526	836,255	868,492	907,691
Females, 60+	477,974	543,943	611,545	687,005	762,453	834,376	905,995	946,632
Couples with children	666,062	685,749	702,135	734,497	767,008	798,916	830,408	861,900
Couples only	482,361	525,238	574,241	614,872	656,979	696,898	735,802	774,707
Single parent	221,754	240,333	257,777	272,580	288,192	304,450	320,780	337,110
Single person	450,884	505,173	564,541	610,453	658,838	709,620	759,916	810,211
Other types	105,477	109,195	113,081	120,932	127,485	134,836	142,509	150,182
Full time workers	1,851,708	1,905,493	1,966,906	2,051,800	2,077,986	2,124,131	2,189,699	2,278,616
Other workers	674,137	745,910	838,565	974,092	1,067,521	1,143,542	1,210,916	1,254,942
$ 0–20,799	1,802,515	1,848,608	1,884,021	1,916,069	1,949,455	1,980,494	2,007,958	2,025,830
$ 20,800–31,199	559,720	581,251	602,371	623,459	645,448	666,676	686,449	702,973
$ 31,200–41,599	459,236	480,995	502,219	522,948	544,714	566,310	587,039	584,929
$ 41,600–67,599	725,312	786,900	850,326	913,710	979,774	1,046,559	1,088,265	1,133,226
$ ≥ 67,600	594,502	713,811	839,457	969,004	1,103,266	1,241,425	1,408,332	1,590,522
Tertiary students	413,005	434,072	456,214	479,486	503,945	529,651	556,668	585,064

Totals by target group

Population	5,132,884	5,440,282	5,755,043	6,076,448	6,401,661	6,724,818	7,041,047	7,357,275
Households	1,926,539	2,065,688	2,211,774	2,353,333	2,498,502	2,644,720	2,789,415	2,934,110
Workers	2,525,845	2,651,403	2,805,471	3,025,892	3,145,507	3,267,674	3,400,615	3,533,557
Adults	4,141,284	4,411,566	4,678,395	4,945,191	5,222,657	5,501,464	5,778,044	6,037,480
Tertiary students	413,005	434,072	456,214	479,486	503,945	529,651	556,668	585,064

Percentage growth relative to 2006 by target

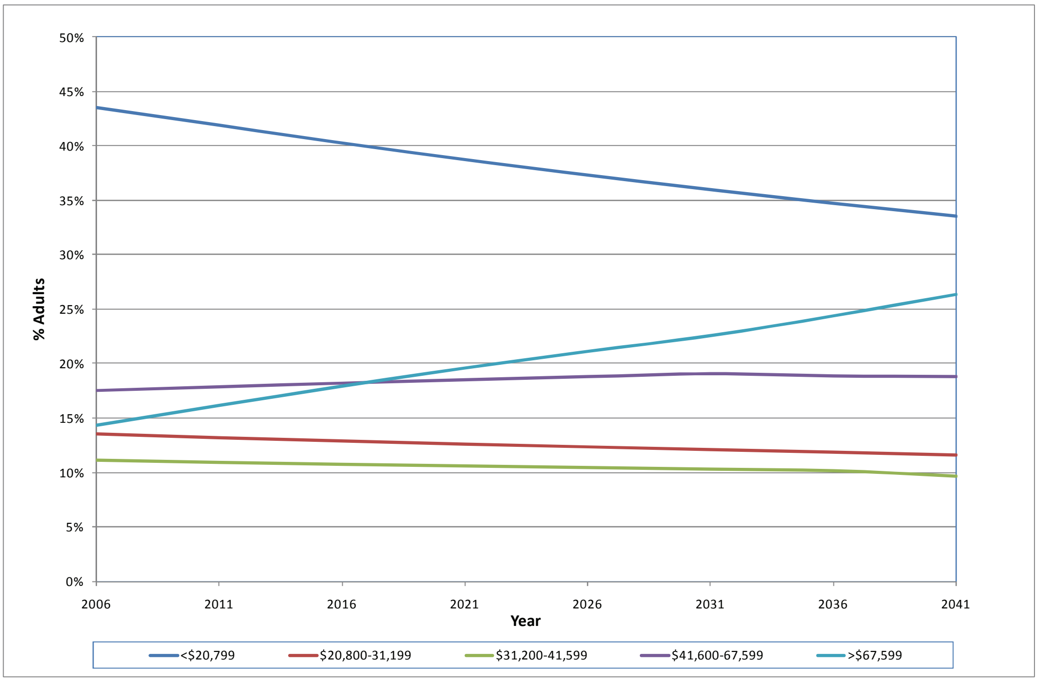
Males, 0-19		4.0%	8.1%	13.0%	18.3%	22.8%	27.1%	32.9%
Males, 20-39		4.5%	9.0%	13.7%	17.4%	21.3%	25.1%	30.7%
Males, 40-59		3.4%	7.6%	10.1%	13.8%	18.6%	23.5%	29.2%
Males, 60+		17.0%	33.4%	51.5%	69.5%	86.1%	102.1%	111.4%
Females, 0-19		3.8%	7.8%	12.5%	17.7%	22.2%	26.4%	32.2%
Females. 20-39		4.7%	8.9%	13.4%	16.7%	20.4%	24.1%	29.4%
Females, 40-59		4.0%	8.5%	11.3%	15.4%	20.3%	24.9%	30.6%
Females, 60+		13.8%	27.9%	43.7%	59.5%	74.6%	89.5%	98.1%
Couples with children		3.0%	5.4%	10.3%	15.2%	19.9%	24.7%	29.4%
Couples only		8.9%	19.0%	27.5%	36.2%	44.5%	52.5%	60.6%
Single parent		8.4%	16.2%	22.9%	30.0%	37.3%	44.7%	52.0%
Single person		12.0%	25.2%	35.4%	46.1%	57.4%	68.5%	79.7%
Other types		3.5%	7.2%	14.7%	20.9%	27.8%	35.1%	42.4%
Full time workers		2.9%	6.2%	10.8%	12.2%	14.7%	18.3%	23.1%
Other workers		10.6%	24.4%	44.5%	58.4%	69.6%	79.6%	86.2%
$ 0–20,799		2.6%	4.5%	6.3%	8.2%	9.9%	11.4%	12.4%
$ 20,800–31,199		3.8%	7.6%	11.4%	15.3%	19.1%	22.6%	25.6%
$ 31,200–41,599		4.7%	9.4%	13.9%	18.6%	23.3%	27.8%	27.4%
$ 41,600–67,599		8.5%	17.2%	26.0%	35.1%	44.3%	50.0%	56.2%
$ ≥ 67,600		20.1%	41.2%	63.0%	85.6%	108.8%	136.9%	167.5%
Tertiary students		5.1%	10.5%	16.1%	22.0%	28.2%	34.8%	41.7%

Percentage growth relative to 2006 by target group

Population		6.0%	12.1%	18.4%	24.7%	31.0%	37.2%	43.3%
Households		7.2%	14.8%	22.2%	29.7%	37.3%	44.8%	52.3%
Workers		5.0%	11.1%	19.8%	24.5%	29.4%	34.6%	39.9%
Adults		6.5%	13.0%	19.4%	26.1%	32.8%	39.5%	45.8%
Tertiary students		5.1%	10.5%	16.1%	22.0%	28.2%	34.8%	41.7%

[5] Note that the income control target for children (income0) is not presented in Table 8, as the income target distribution has been calculated for adults only.

Figure 8: Changes in the proportion of adults in each income band over time



An important point to note is that the simple assumption of 1% growth in real terms income per annum used to derive the future income targets does not take account of demographic impacts. In particular, as Sydney's population is ageing, the proportion of persons of working age is forecast to decrease over time, and this would have the impact of reducing the real terms growth in incomes.

3.3 Treatment of income

This section starts by summarising how incomes were treated in the 2001 base version of the Population Synthesiser, created in 2004, and used in the 2001 base version of the Sydney Strategic Model (STM).

Section 3.3.2 describes developments in RAND Europe's thinking regarding forecasting incomes since the 2001 base version of the Population Synthesiser was developed, specifically a decomposition of income growth into welfare and redistribution components.

Section 3.3.3 details how the separate welfare and redistribution components have been incorporated into the new 2006 base Population Synthesiser.

3.3.1 Treatment in current STM (2001 base)

The treatment of incomes in the 2001 base STM was documented in full in Fox *et al* (2004). An important point to note is that while the 2001 base version of the STM had a

2001 base, the base year for prices in the demand models was 1996, and therefore the 2001 base version of the Population Synthesiser works in 1996 prices[6].

Table 9 summarises how income growth was treated in each of the Population Synthesiser components.

Table 9: Treatment of income in 2001 base Population Synthesiser

Component	Income Measure	Income Treatment
Licence Models	Household	Income = Income * Inc_Factor
Car Ownership Models	Household	Income = Income * Inc_Factor
Prototypical Sample	Personal Income	Re-distribution between four targets
Category Segments	Personal and household	Income = Income * Inc_Factor
Accumulation	Personal and household	Income = Income * Inc_Factor

The 'inc_factor' defined the real increase in income for future year runs relative to 2001 levels. So a factor of 1.1 implied a 10% real increase relative to 2001. The same uniform increase in income was assumed for all individual and household incomes, and this uniform increase in income resulted in higher predicted licence holding and car ownership levels in future years. It was necessary for the user to manually specify the income factor as an input into the process on the basis of overall forecast real terms growth in income relative to the base year. An assumption of a 1% real terms growth in incomes per annum was used.

In the prototypical sampling, re-distribution between income bands occurred as the sample was expanded to meet four specific income targets, as well as 15 other targets, which included two worker targets.

The definition of income bands was complicated by the fact that the 2001 Census used bands defined in 2001 prices, which needed to be deflated to 1996 prices. The income targets in 1996 prices were, to the nearest $ 10 per annum:

- $ 0–5,529
- $ 5,530–18,339
- $ 18,440–36,879
- $ 36,880 +

It is noted that both the Census and the HTS record incomes using bands rather than the exact income level. Furthermore, both the Census and the HTS use more than four bands, so that there is an aggregation to the four bands detailed above.

A procedure was developed to take account of missing incomes in the 2001 Census data, based on the assumption that missing incomes had the same income distribution as stated incomes. Before the population synthesiser was applied for forecast years, a procedure was applied to take account of real terms income growth that shifted a proportion of the target population from lower to higher income bands. The population synthesiser was then run using these adjusted targets.

6 Specifically, deflated to 1996 Census Journey to Work prices levels, using CPI = 120.1.

When the prototypical sampling was applied, the sample was expanded by weighting across household categories. These household categories did not include income, but they did include the number of workers in the household, which is closely correlated with household income.

The dimensions over which the weighting takes place are:

- the number of adults in the household (1, 2, 3, 4 or more);
- the presence of children (yes or no);
- the number of workers in the household (0, 1, 2, 3, 4 or more); and
- age of head of household (under 30, 30–45, 45–65, over 65).

This categorisation yields 160 categories. However, because of logical conditions (workers cannot exceed adults) and the merging of small categories, the final number of household categories used in the expansion of the 2001 base version of the Population Synthesiser is 74.

3.3.2 Decomposition of income growth

Since the 2001 base Population Synthesiser was developed, RAND Europe has further developed its thinking on the treatment of income growth in Population Synthesisers of the type used in the STM.

Experience from developing the PRISM model for the West Midlands region of the UK[7] indicated that if income is specified as a target, increases in incomes relative to the base year can best be achieved by giving higher weights to multi-worker households in the future. This can result in an over-prediction of the number of workers relative to the zonal targets.

The solution that was developed for the PRISM model and that has been used in the STM model is to de-compose income growth into two components:

- a 'welfare increase', a uniform increase applied to the incomes of all persons; and
- redistribution, where some individuals move from one income band to another.

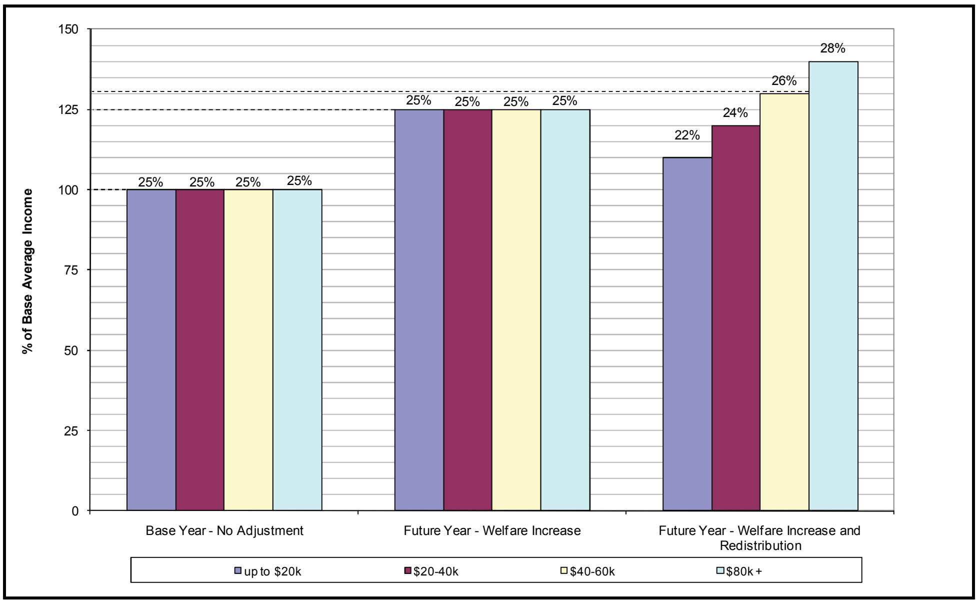
The welfare increase is applied to explain the real terms increase in average pay per worker (and benefit levels). The redistribution effect accounts for further change in incomes due to individuals shifting between income bands. Changes in general welfare and shifts between categories are independent, because the categories are not defined by income.

This process is illustrated in Figure 9. In this example, a welfare increase of 25% is applied to all incomes, and then further growth in incomes is achieved through re-distribution from lower to higher income bands.

---

[7] See http://www.prism-wm.com/ for general information on PRISM. There is currently no published paper on RAND Europe's treatment of incomes in Population Synthesisers.

Figure 9: Decomposition of income growth



In Figure 9, in the base year, 25% of the population lies in each income band, and by definition the average income is 100% of the base year value.

When the welfare increase is applied, all incomes are increased by a uniform 25%, and there is no shifting between bands, so each band still contains 25% of the population. Therefore, the average income is 125% of the base year value.

The final step is to apply some re-distribution between bands, with some individuals shifting from lower income to higher income bands. The net result of these shifts, based on simple assumptions about the average income in each band, is that overall average income is 131.25% of the base year value.

The appropriate balance between the welfare increase and the re-distribution effect is identified using an iterative procedure, where the process is run with a series of different assumptions, and the fit to the income and workers target is assessed. RAND Europe's experience from the PRISM model, and a model of long-distance travel in the UK, is that the welfare increase is the stronger effect. However, the exact balance between the two effects would be expected to vary between contexts.

The approach used in the 2001 base Population Synthesiser effectively applied a total income increase rather than a welfare increase, and the welfare and redistribution effects of overall income growth were not reconciled. Section 3.3.3 details how these two effects have been better reconciled in the 2006 base Population Synthesiser.

3.3.3 Treatment in new STM (2006 base)

The Population Synthesiser has been modified to use the welfare and re-distribution approach. This required only relatively minor changes relative to how the 2001 base system operated.

The approach is summarised in Table 10.

Table 10: Treatment of income in 2006 base Population Synthesiser

Component	Income Measure	Income Treatment
Licence Models	Household	Income = Income * Welfare Factor
Car Ownership Models	Household	Income = Income * Welfare Factor
Prototypical Sample	Personal Income	Re-distribution between four targets
Category Segments	Personal and household	Income = Income * Welfare Factor
Accumulation	Personal and household	Income = Income * Welfare Factor

The welfare factor replaces 'inc_fact' in the 2001 base system, as now only a proportion of the total income growth is applied through the welfare factor, and the remainder of income growth is achieved using re-distribution. Note that at the ACCUM stage, the outputs from the category segment process, which take account of the welfare increase, are combined with the QUAD expansion factors from the prototypical sampling procedure, which take account of re-distribution. Thus, at the ACCUM stage the two income types of income growth are brought together.

For a given forecast year, the appropriate value for the welfare factor and re-distribution has been investigated by running the prototypical sampling procedure iteratively, and assessing the match across the target variables, and in particular across worker and income targets. These runs are documented in Section 5.2.1.

Five income targets have been used (defined in 2006 prices):

- < $ 20,800
- $ 20,800–31,199
- $ 31,200–41,599
- $ 41,600–67,599
- > $ 67,599

These income bands are consistent with the five bands used in the home–work income segmentation, which are used to segment both willingness to pay and the employment attraction variable.

An issue that arose for model application was that the incomes recorded in the prototypical sample were mid-point incomes for the 11 personal income bands recorded in the HTS data. When welfare factors are applied to increase these incomes for future years, at a certain welfare factor threshold, a group of individuals suddenly move up from one income band to another. This feature resulted in lumpy shifts between income bands, which resulted in a poor fit to the income targets for certain years.

To overcome this issue, incomes were imputed by drawing a random variable from a uniform (0,1) distribution, and applying this to impute an income within each of the bands used in the HTS. This process operates as follows:

$$inc_i = inc_L + U(0,1) * \left(inc_U - inc_L\right)$$

where: inc_i is the imputed income for an individual

$U(0,1)$ is a random variable drawn from a uniform (0,1) distribution

inc_L is the lower band for the individual's income

inc_U is the upper band for the individual's income

For individuals with zero recorded income, no income is imputed as they are treated as having zero income for all years. For individuals with incomes greater than $ 67,599 in the base, there is no need to impute an income, as they will always be in the top income band (it is assumed that income growth is always positive relative to the 2006 base).

3.4 Area types

During the re-estimation of the mode-destination models (Fox *et al*, 2010), a number of area type terms were added to the models to reflect differences in mode usage between inner, middle and outer Sydney rings, Newcastle and Wollongong, not explained by personal, household or service characteristics.

Furthermore, when the 2001 base version of the Population Synthesiser was validated in 2004 (Fox *et al*, 2004), a general theme of the validation was that the model predictions were not always able to predict the spatial variation in the observed data.

Therefore, investigations were undertaken to investigate whether the Population Synthesiser should be run separately by four area types:

1. inner Sydney ring;
2. middle Sydney ring;
3. outer Sydney ring; and
4. Newcastle and Wollongong combined (regional).

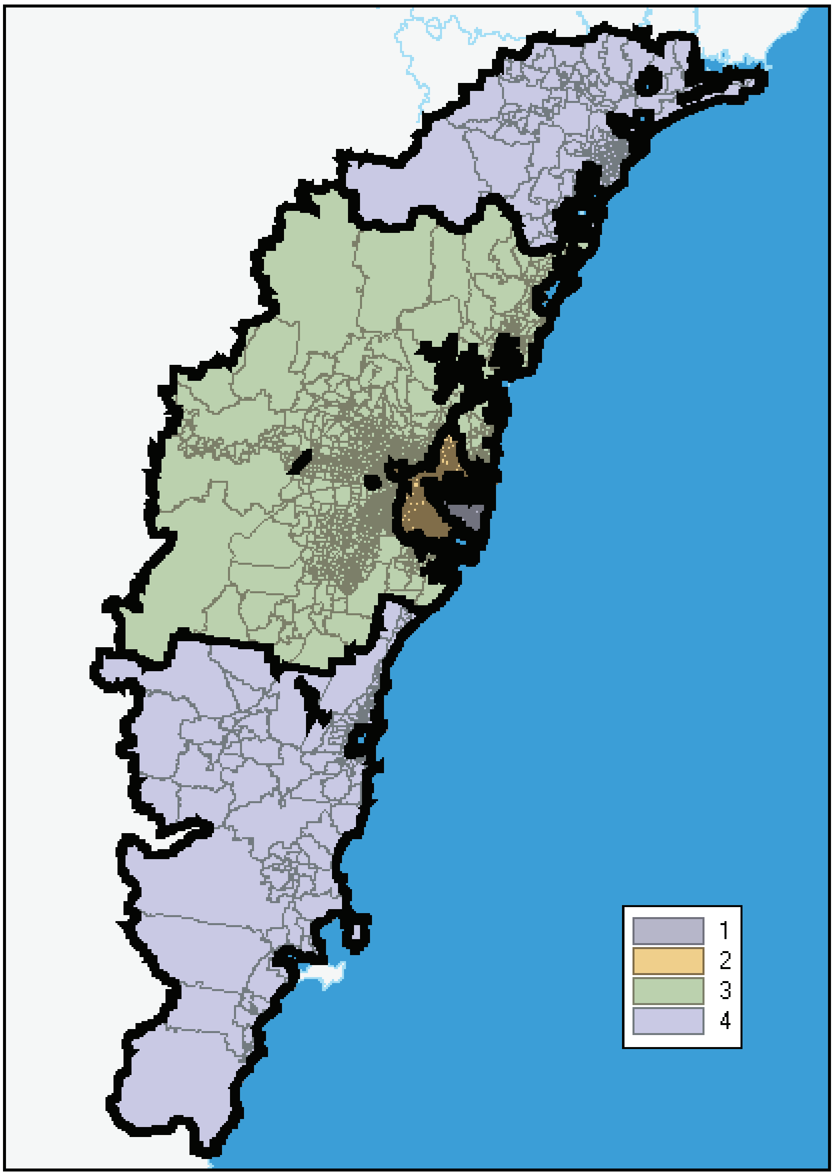
It is noted that the 2001 base version of the Population Synthesiser produced forecasts for area types 1 to 3 only; the extension to cover Newcastle and Wollongong is an addition to the 2006 base version of the Population Synthesiser.

The unique level of geography which defines a ring is the 2006 statistical local area (SLA). Appendix B provides a full definition of the four area types, listing:

- the statistical local area (SLA)
- the statistical division (SD)
- the statistical sub-division (SSD)

Figure 10 plots the four area types across the whole study area, and then Figure 11 zooms in to better show the area covered by the inner and middle Sydney rings.

Figure 10: Area types across the study area



Note that area type 4 comprises both an area to the north that covers Newcastle and an area to the south that covers Wollongong.

Figure 11: Inner and middle area types

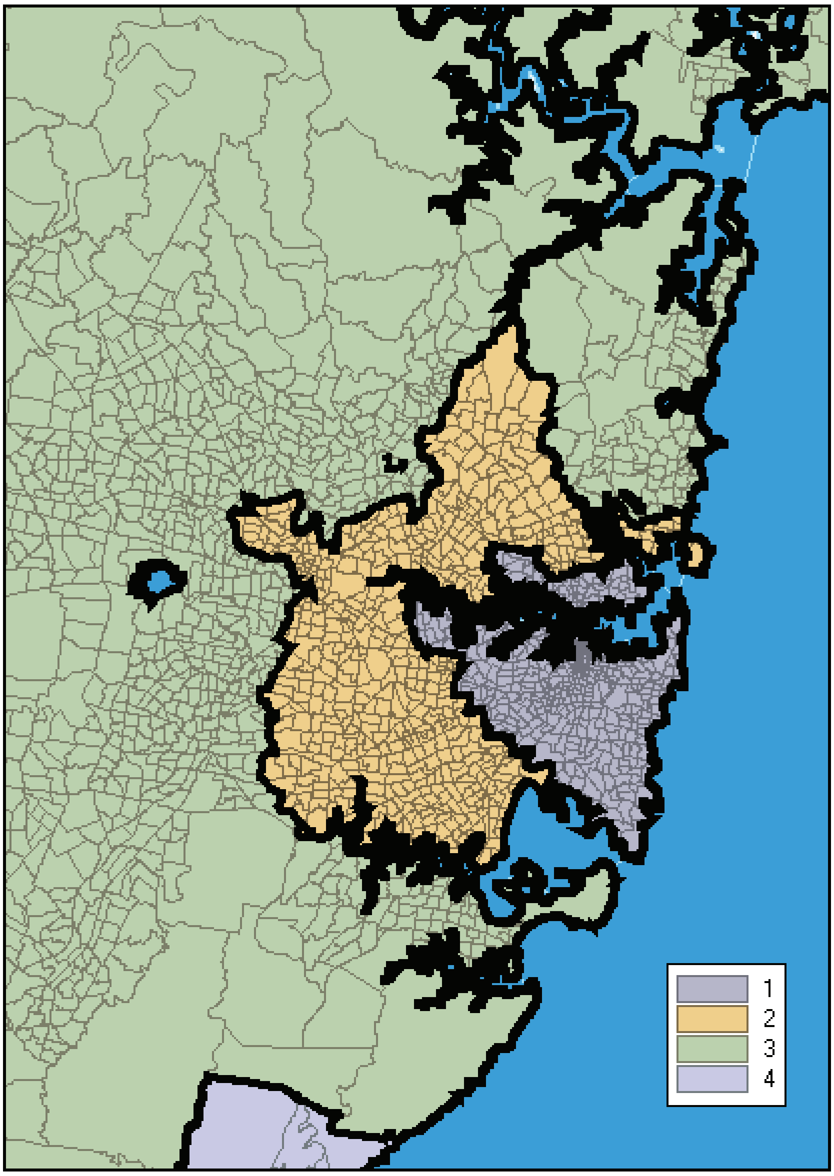


Table 11 summarises the population of each area type by age band for the 2006 base year.

Table 11: Population by age and area type, 2006

	Inner	Middle	Outer	Regional	Total
0–14	100,706	213,910	495,248	179,955	989,820
15–49	470,131	584,704	1,154,854	426,679	2,636,368
50+	214,628	335,986	647,714	311,558	1,509,886
Total	785,465	1,134,601	2,297,816	918,192	5,136,074

	Inner	Middle	Outer	Regional	Total
0–14	12.8 %	18.9 %	21.6 %	19.6 %	19.3 %
15–49	59.9 %	51.5 %	50.3 %	46.5 %	51.3 %
50+	27.3 %	29.6 %	28.2 %	33.9 %	29.4 %
Total	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %

The area with the largest population is the outer Sydney ring, which has more than double the population of the middle Sydney ring, the next largest area. The inner Sydney ring has the lowest population.

In terms of age profile, the inner ring has the lowest percentage of persons aged 0–14, and the highest percentage of persons aged 15–49, as would be expected for an area with higher than average numbers of young professional residents and students. The highest percentage of children aged 0–14 is found in the outer ring, which is more suburban in nature than the inner and middle rings.

Table 12: Households by household type and area type, 2006

	Inner	Middle	Outer	Regional	Total
Single	118,687	94,751	157,071	89,299	459,808
Couple	92,776	94,537	186,901	99,045	473,258
Couple with Children	78,856	158,754	331,169	112,520	681,298
Single with Children	27,819	45,800	102,986	44,169	220,774
Other (Group)	40,779	23,089	28,859	14,796	107,523
Total	358,916	416,931	806,985	359,829	1,942,662

	Inner	Middle	Outer	Regional	Total
Single	33.1 %	22.7 %	19.5 %	24.8 %	23.7 %
Couple	25.8 %	22.7 %	23.2 %	27.5 %	24.4 %
Couple with children	22.0 %	38.1 %	41.0 %	31.3 %	35.1 %
Single with children	7.8 %	11.0 %	12.8 %	12.3 %	11.4 %
Other (Group)	11.4 %	5.5 %	3.6 %	4.1 %	5.5 %
Total	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %

Substantial differences in the household type distributions can be observed between area types. Inner Sydney has the highest percentage of single households and other (group) households, and the lowest percentage of households with children. Outer Sydney has the highest percentage of households with children, which is consistent with its more suburban development pattern.

In summary, there are significant differences between the age profiles, and between the household type profiles, across the four area types. Furthermore, it is straightforward to run the QUAD process separately by area type. Therefore, separate QUAD runs were made by the four area types in the 2006 base Population Synthesiser. Section 5.1.4 presents the results of base year testing. It demonstrates that this approach allows a better fit to the Census target data than making a single study area wide run.

3.5 Household categories

This section summarises the changes to the household categories in the Population Synthesiser. The number of household categories has been reduced relative to the 2001 base version of the Population Synthesiser because when the household sample is split by the four area types discussed in Section 3.4, the number of households in less usual categories is lower, and so more merging of categories is necessary.

The 74 household categories in the 2001 base version of the Population Synthesiser is the starting point. A minimum cell value of 6 is achieved by aggregating the 74 household categories to 59. The value 6 was chosen to achieve a reasonable balance between the reliability of the distribution and potential loss of detail.

To move from the previous categorisation (CAT) to the new categorisation (CAT2) 12 aggregation steps were followed:

CAT 1 and 3 into CAT2 1: households with the household head aged between 18 and 29, only a single adult, and no worker were originally separated into two categories – with and without children (CAT 1 and 3 in the 2001 base Population Synthesiser). These are merged into one (CAT2 1).

CAT 2 and 4 into CAT2 2: households with the household head aged between 18 and 29, only a single adult, and one worker were originally separated into two categories – with and without children (CAT 2 and 4 in the 2001 base Population Synthesiser). These are merged into one (CAT2 2).

CAT 15 and 18 into CAT2 13: households with the household head aged between 18 and 29, two adults, and no workers were originally separated into two categories – with and without children (CAT 15 and 18 in the 2001 base Population Synthesiser). These are merged into one (CAT2 13).

CAT 21 and 24 into CAT2 18: households with the household head aged between 30 and 44, two adults, and no workers were originally separated into two categories – with and without children (CAT 21 and 24 in the 2001 base Population Synthesiser). These are merged into one (CAT2 18).

CAT 27 and 30 into CAT2 23: households with the household head aged between 45 and 64, two adults, and no workers were originally separated into two categories – with and without children (CAT 27 and 30 in the 2001 base Population Synthesiser). These are merged into one (CAT2 23).

CAT 39, 40 and 41 into CAT2 34: households with the household head aged between 18 and 29, three adults, and children were originally separated into three categories – with one, two or three workers (CAT 39, 40, and 41 in the 2001 base Population Synthesiser). These are merged into one (CAT2 34).

CAT 54, 55, 56 into CAT2 47: households with the household head aged 65 or over, three adults, and with or without children were originally separated into three categories – with one, two or three workers (CAT 54, 55, and 56 in the 2001 base Population Synthesiser). These three are merged into one (CAT2 47).

CAT 58 and 59 into CAT2 49: households with the household head aged between 18 and 29, four or more adults, and no children were originally separated into three categories – with two, three or four+ workers (CAT 57, 58, and 59 in the 2001 base Population Synthesiser). Those with 3 or 4+ workers (CAT 58 and 59) are merged into one (CAT2 49). Category 57 is not merged, only renumbered to CAT2 48.

CAT 60, 61 and 62 into CAT2 50: Households with the household head aged between 18 and 29, four or more adults, and children were originally separated into three categories – with two, three or four+ workers (CAT 60, 61, and 62 in the 2001 base Population Synthesiser). These three are merged into one (CAT2 50).

CAT 63, 64 and 65 into CAT2 51: Households with the household head aged between 30 and 44, four or more adults, and no children were originally separated into three categories – with two, three or four+ workers (CAT 60, 61, and 62 in the 2001 base Population Synthesiser). These three are merged into one (CAT2 51).

CAT 67 and 68 into CAT2 53: households with the household head aged between 30 and 44, four or more adults, and children were originally separated into two categories – with three or four+ workers (CAT 67, and 68 in the 2001 base Population Synthesiser). These two are merged into one (CAT2 53).

Table 13 summarises the new household categories CAT2, with a definition of each category and the correspondence to the old household categories CAT used in the 2001 base Population Synthesiser.

Table 13: Summary of the revised household categories CAT2

CAT2	Age of Head	Adults	Children?	Workers	CAT
1	18–29	1	yes or no	none	1, 3
2	18–29	1	yes or no	1	2,4
3	30–44	1	no	none	5
4	30–44	1	no	1	6
5	30–44	1	yes	none	7
6	30–44	1	yes	1	8
7	45–64	1	no	none	9
8	45–64	1	no	1	10
9	45–64	1	yes	none	11
10	45–64	1	yes or no	1	12
11	65+	1	yes or no	none	13
12	65+	1	do not care	1	14
13	18–29	2	yes or no	none	15,18
14	18–29	2	no	1	16
15	18–29	2	no	2	17
16	18–29	2	yes	1	19
17	18–29	2	yes	2	20
18	30–44	2	yes or no	none	21,24
19	30–44	2	no	1	22
20	30–44	2	no	2	23
21	30–44	2	yes	1	25
22	30–44	2	yes	2	26
23	45–64	2	yes or no	none	27,30
24	45–64	2	no	1	28
25	45–64	2	no	2	29
26	45–64	2	yes	1	31
27	45–64	2	yes	2	32
28	65+	2	yes or no	none	33
29	65+	2	yes or no	1	34
30	65+	2	yes or no	2	35
31	18–29	3	no	1	36
32	18–29	3	no	2	37
33	18–29	3	no	3	38
34	18–29	3	yes	1,2 or 3	39,40,41

CAT2	Age of Head	Adults	Children?	Workers	CAT
35	30–44	3	no	1	42
36	30–44	3	no	2	43
37	30–44	3	no	3	44
38	30–44	3	yes	1	45
39	30–44	3	yes	2	46
40	30–44	3	yes	3	47
41	45–64	3	no	1	48
42	45–64	3	no	2	49
43	45–64	3	no	3	50
44	45–64	3	yes	1	51
45	45–64	3	yes	2	52
46	45–64	3	yes or no	3	53
47	65+	3	do not care	1, 2, or 3	54, 55, 56
48	18–29	4+	no	2	57
49	18–29	4+	no	3+	58, 59
50	18–29	4+	yes	2+	60, 61, 62
51	30–44	4+	no	2+	63,64,65
52	30–44	4+	yes	2	66
53	30–44	4+	yes	3+	67,68
54	45–64	4+	no	2	69
55	45–64	4+	no	3	70
56	45–64	4+	no	4+	71
57	45–64	4+	yes	2	72
58	45–64	4+	yes	3	73
59	45–64	4+	yes	4+	74

CHAPTER 4 Car ownership pivot

The 2004 STM validation project (Fox *et al*, 2004) revealed that while average car ownership levels across the STM study area were predicted well, the model was not able to fully re-produce the spatial variation in car ownership levels observed in the 2001 Census. Therefore, a new component has been added to the 2006 base Population Synthesiser to 'pivot' off the car ownership levels observed for each model zone in the 2006 Census.

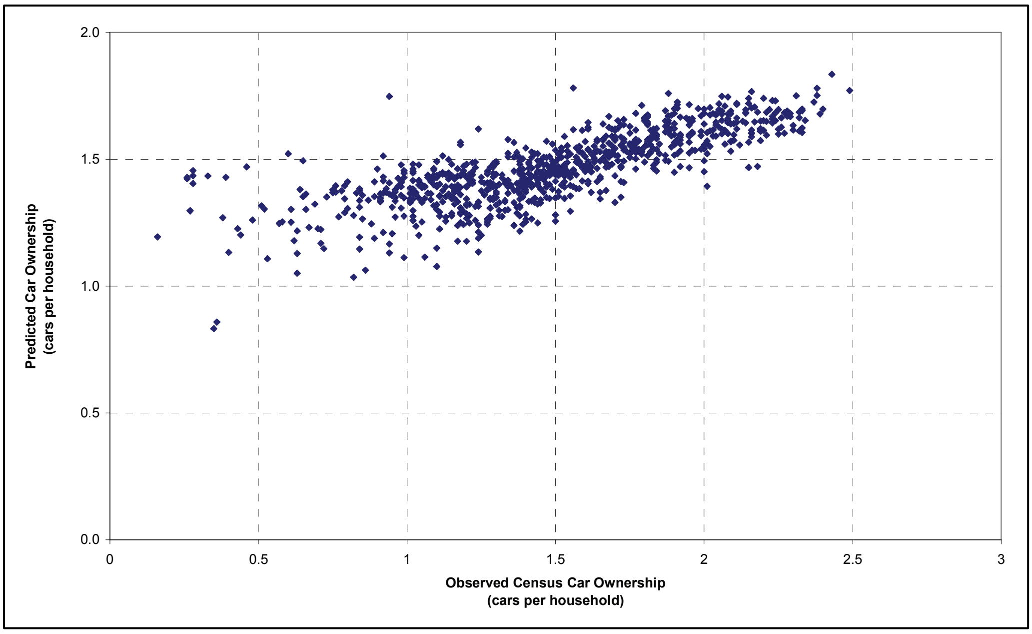
Section 4.1 provides more details about the findings of the 2004 validation project that forms the background to this work. Section 4.2 introduces the pivot procedure that has been used for this study, and sets of the special procedure used to estimate the car ownership pivot model. Section 4.3 then describes the Census data that provides information on zonal car ownership levels in the 2006 base, and in particular describes the procedure followed to ensure that the observed car ownership levels were reliable for low population zones. Section 4.4 defines the various segmentations that are required to implement the car ownership pivot. Finally, building upon the procedures set out in Section 4.2, more detail on the model estimation and calibration steps is given in Sections 4.5 and 4.6 respectively.

4.1 Background

During the 2004 validation project (Fox *et al*, 2004) the predictive performance of the car ownership model was assessed across different spatial areas. The analysis compared car ownership levels observed in the 2001 Census to the predictions of the 2001 base version of the total car ownership model.

The conclusions from the spatial validation were that while the model predicted observed variation in ownership between Sydney Statistical Divisions (SSDs) well, at the zonal level the models were not able to fully replicate the observed variation in car ownership levels. This is illustrated in Figure 12, a zonal scatter plot reproduced from Fox *et al* (2004).

Figure 12: Comparison of observed and predicted car ownership by zone (2001 base model)



In this figure, it can be seen that the spread in the observed car ownership levels (the x-axis) was greater than the spread in the predicted car ownership levels (the y-axis). It should be emphasised that there was a strong positive correlation between observed and predicted ownership levels of 0.79. Therefore, the models did successfully predict higher levels of car ownership where observed levels were higher; the issue was that the full range of values was not replicated by the model predictions.

The impact of these spatial differences was that while overall car ownership levels feeding into the travel demand models were accurate for 2001 in comparison with Census data, the spatial differences in car ownership were not fully replicated. Therefore, in areas where car ownership was substantially lower than average, ownership levels closer to the mean were predicted, which in turn resulted in over-prediction of the car driver share for these areas. Similarly, in areas where car ownership was noticeably higher than average, predicted ownership levels were again closer to the overall mean, which in turn resulted in some under-prediction of the car driver share for these areas.

By introducing a car ownership pivot procedure to the Population Synthesiser, the full spatial variation in car ownership observed in the 2006 base year is better reproduced by the model, which in turn ensures the model is better able to predict spatial variations in the car driver share. For future years, the same set of zonal correction factors is applied. This is analogous to other forms of pivoting where the differences between base year predicted and observed values – sometimes termed K-factors – are carried forward from the base to future years. So, in essence, the calibration of the pivot procedure is one of identifying the zonal corrections to be applied.

4.2 Pivot procedure

The total car ownership model represents the choice between four alternatives:

- zero cars
- one car
- two cars
- three-plus cars

Consideration was given to specifying the observed car ownership levels from the Census for each of these four alternatives for each of the 2,690 model zones. However, the percentages of households with zero and three-plus car households are low for some area types, as illustrated in Table 14. It is noted that the Census data has separate categories for three and four-plus vehicle households; these have been combined in Table 14.

Table 14: Car ownership state percentages by area type (2006 Census data)

Area type	0 Cars	1 Car	2 Cars	3+ Cars	Total
Inner	24.0 %	47.0 %	23.3 %	5.6 %	100.0 %
Middle	14.9 %	42.4 %	31.5 %	11.1 %	100.0 %
Outer	9.8 %	36.4 %	37.8 %	16.0 %	100.0 %
Regional	10.7 %	39.8 %	35.7 %	13.9 %	100.0 %
Total	13.7 %	40.3 %	33.4 %	12.7 %	100.0 %

Only 5% of households in the inner Sydney ring have three or more cars, and in the outer Sydney ring under 10% of households have zero cars.

Therefore a single 'target' for each model zone, namely total cars per household, was defined. This allows a more straightforward calibration procedure, with a single correction parameter estimated for each model zone, and requires less information from the zonal car ownership data, in particular for less frequently chosen car ownership states.

It was also decided in discussion with the BTS that company car ownership would not be calibrated at the spatial level. The reasons for not doing this are first that the total car ownership model is applied after the company car models, so that calibrating to total cars allows overall car ownership levels to be matched exactly at the zonal level; and second that the Census data details total car ownership only, so expanded Household Travel Survey data would have to be used to define base year company car ownership levels by area, and this would not support a company car ownership pivot at the zonal level because of the sample sizes (28,434 households for 2,690 zones gives 10.6 households per zone on average).

The pivot procedure has been set up as an incremental logit model applied relative to the (uncalibrated) model probabilities:

$$p_{iz}^{*} = \frac{p_{iz}\exp(\beta_z c_i)}{\sum_j p_{jz}\exp(\beta_z c_j)} \qquad (4.1)$$

where: p_{iz}^{*} is the probability of car ownership level i for zone z after pivoting

p_{iz} is the probability of car ownership level i for zone z before pivoting (from the total car ownership model TOTCAR)

i is one of the j possible car ownership levels (0, 1, 2, 3+ cars)

z is the model zone

c_i is the number of cars owned at the car ownership level

β_z is the zonal calibration constant

It is noted that the number of cars owned at each ownership level is simply $c_j=j$, except for $j=3$ where the mean number of cars for three or more car households of 3.4356 was used[8]. It is necessary to consider each of the four car owning alternatives in order to determine the overall level of car ownership predicted by the total car ownership model.

In order to estimate Equation 4.1, a special *sufficient statistics* run was made. A sufficient statistics run is a special type of estimation run that allows new parameters to be estimated using a disaggregate logit model, such as the total household cars model, with choice defined as an aggregate proportion, such as cars per adult. The ALOGIT software has a special procedure that allows runs of this type to be made.

To explain how this procedure works, it is helpful to quote Section 3.7 of Train (2003)[9], who states the first order conditions for estimation of each of the parameters β of a logit model as:

$$\sum_n \sum_i y_{ni} x_{ni} = \sum_n \sum_i P_{ni} x_{ni} \qquad (4.2)$$

where: n is the decision maker

i is the alternative

y_{ni} is 1 if the alternative is chosen, 0 otherwise

P_{ni} is the predicted probability for ni

x_{ni} is the value of the attribute for ni

Equation 4.2 gives the condition that the observed proportion of each variable x for which a parameter β is estimated is equal to the predicted proportion.

Introducing weights to this expression, Equation 4.2 becomes:

$$\sum_n \sum_i w_n y_{ni} x_{ni} = \sum_n \sum_i w_n P_{ni} x_{ni} \qquad (4.3)$$

where: w_n is the weight associated with each decision maker

In the sufficient statistics procedure, we can take advantage of the fact that the only parameter to be estimated is β_z, the zonal correction factor, and the observed choice proportion we want to achieve for a given zone is the target cars per household. This allows us to write:

$$T = \sum_n \sum_i w_n P_{ni} c_i \qquad (4.4)$$

8 The mean is calculated from unweighted HTS data collected between July 1999 and June 2008 to maximise the sample size. If the mean is calculated from unweighted HTS data collected between July 2004 and June 2008 (the four waves used to define the Prototypical Sample) a slightly higher value of 3.4688 is obtained, which is 0.0322 cars per household higher.

9 There is a much earlier article on estimating logit choice models by Daly (1982).

where: T is the target cars for the zone

n is the household

i is the set of car ownership alternatives

w_n is the weight for the household

P_{ni} is the probability of each car ownership state

c_i is the number of cars associated with each car ownership state

Section 4.5 provides further detail explaining how Equation 4.4 has been estimated for this specific application, and Section 4.5 details how the car ownership model has been implemented as part of the overall Prototypical Sampling procedure.

4.3 Census data

Data from the 2006 Census has been assembled for each of the 2006 model zones to define observed base year car ownership levels by zone for the calibration of the car ownership pivot model. For each zone, the data reports the number of households with zero, one, two, three and four-plus cars. The distribution of the Census data across these five car ownership categories is compared to the 2004-08 HTS data in Table 15, in both cases using data from across the model study area.

Table 15: Comparison of Census (2006) and HTS (2004-08) car ownership levels

	Census	HTS (weighted)	Difference
0 Cars	13.3 %	13.7 %	-0.4 %
1 Car	39.9 %	40.8 %	-0.9 %
2 Cars	33.8 %	33.4 %	0.4 %
3 Cars	9.0 %	8.3 %	0.7 %
4+ Cars	4.0 %	3.8 %	0.2 %
Total	100.0 %	100.0 %	0.0 %

Table 15 demonstrates that there is a close match between the Census and weighted HTS car ownership distributions across the study area.

The Census data was used to obtain data on the average number of cars per household, as discussed below. A procedure was devised by the BTS to deal with zones with low numbers of households to ensure that the car ownership information was reliable:

1. For each STM zone, the number of households that stated their car ownership level was determined.
2. For zones with 90 or more households, the distribution between zero, one, two and three-plus car households was taken directly from the Census data.
3. For zones with fewer than 90 households:
    a. the BTS ‘borrowed’ households from the nearest neighbouring zone with at least 90 households;
    b. the original and borrowed households were added together and used to determine the proportion of households with zero, one, two and three-plus households; and

   c. these proportions were applied to the original number of households in the zone (always fewer than 90) to determine the number of households with zero, one, two and three-plus cars.

4. For each household, the one, two, three and four-plus car[10] proportions were used to calculate the mean number of cars per household (for low population zones this is the original number of households prior to 'borrowing' households from neighbouring zones).

Car ownership information was borrowed from a neighbouring zone for 515 (19.1%) of the 2,690 model zones. However, by definition these zones contain fewer households than average, so that just 0.6% of households reside in these 515 zones.

At the end of this process, the car ownership targets for each model zone were defined (T in Equation 4.4).

4.4 Car availability segmentations

The starting point for the car availability segmentation to be used in the pivot process is the extended car availability segmentation *aext2*. This segmentation was specified during the model development work undertaken between 2000 and 2001 (HCG and ITS, 2002) and was defined in order to satisfy two requirements:

- It defines all of the mode-destination segmentations used in the home-based travel demand models, defining relevant combinations of individual and household licence holding, company car ownership at the household level, and total car ownership at the household level.
- It allows shifts between groups of segments that can occur as a result of changes in total car ownership that result from changes in accessibility.

The 16 *aext2* segments are defined in Table 16. Competition for cars means that the number of household licences is greater than the number of cars in the household. Free car use means that the number of household licences is less than or equal to the number of cars in the household.

10 It was assumed that the mean number of cars owned for four-plus car households was exactly four following discussions with the BTS.

Table 16: Extended car availability segmentation *aext2*

aext2	Definition
1	No cars, no licence in household
2	No cars, personal licence, one licence in household
3	No cars, no personal licence, at least one licence in household
4	No cars, personal licence, two plus licences in household
5	No personal licence, one non-company car
6	Competition for car, one non-company car
7	Free car use, one household licence, one plus non-company car[11]
8	Competition for car, one company car
9	Free car use, one household licence, one plus company car[11]
10	No personal licence, one company car
11	No personal licence, two plus non-company cars
12	Competition for cars, two plus non-company cars
13	Free car use, several household licences, no company car
14	Competition for cars, one plus c.car, two plus cars in total
15	Free car use, several household licences, one plus company car
16	No personal licence, one plus c.car, two plus cars in total

As illustrated in Figure 2, the total car ownership model is applied after the licence holding and company car ownership models. Therefore, it is necessary to define segments *h*, which give the possible combinations of company car ownership, individual licence and other licence holding within which shifts between the *j* car ownership states can take place in order to meet the target cars per adult. These *h* segments are defined in Table 17. It should be noted that for a household to own a company car it must contain at least one licence holder.

Table 17: Car ownership pivot segments *h*

Segment h	Company Cars	Individual Licence	Other Household Licences
1	0	No	0
2	0	No	1+
3	0	Yes	0
4	0	Yes	1
5	0	Yes	2+
6	1	No	1+
7	1	Yes	0
8	1	Yes	1
9	1	Yes	2+
10	2+	No	1+
11	2+	Yes	0
12	2+	Yes	1
13	2+	Yes	2+

The possible combinations of the extended car availability segmentation *aext2* and the car ownership pivot segments *h* then define the (further) extended car availability segmentation *aext3*. The 37 possible *aext3* segments are defined in Table 18.

[11] It is assumed that for these segments, multiple car dummies are not applicable, as there is only a single licence holder to drive the cars.

Table 18: *aext3* definition and mapping to *aext2* and *h*

aext3	Tot Cars	Comp Cars	Indiv Lic	Other Lics	Car Comp	*aext2*	h
1	0	0	0	0	0	1	1
2	0	0	1	0	0	2	3
3	1	0	1	0	0	7	3
4	1	1	1	0	0	9	7
5	2	0	1	0	0	7	3
6	2	1	1	0	0	9	7
7	2	2+	1	0	0	9	11
8	3+	0	1	0	0	7	3
9	3+	1	1	0	0	9	7
10	3+	2+	1	0	0	9	11
11	0	0	0	1+	0	3	2
12	1	0	0	1+	?	5	2
13	1	1	0	1+	1	10	6
14	2	0	0	1+	?	11	2
15	2	1	0	1+	?	16	6
16	2	2	0	1+	?	16	10
17	3+	0	0	1+	?	11	2
18	3+	1	0	1+	?	16	6
19	3+	2+	0	1+	?	16	10
20	0	0	1	1	0	4	4
21	1	0	1	1	1	6	4
22	1	1	1	1	1	8	8
23	2	0	1	1	0	13	4
24	2	1	1	1	0	15	8
25	2	2+	1	1	0	15	12
26	3+	0	1	1	0	13	4
27	3+	1	1	1	0	15	8
28	3+	2+	1	1	0	15	12
29	0	0	1	2+	0	4	5
30	1	0	1	2+	1	6	5
31	1	1	1	2+	1	8	9
32	2	0	1	2+	1	12	5
33	2	1	1	2+	1	14	9
34	2	2+	1	2+	1	14	13
35	3+	0	1	2+	0	13	5
36	3+	1	1	2+	0	15	9
37	3+	2+	1	2+	0	15	13

For *aext3* segments 35, 36 and 37, it is assumed that the household has free car use, i.e. that the number of household licences is great or equal than the total number of cars.

Table 19 summarises the *aext3* codes that correspond to each *h* segment. This defines the groups within which shifting between segments takes place in order to meet the cars per adult target for the zone.

Table 19: *aext3* codes for each car ownership pivot segment *h*

Segment h	aext3
1	1
2	11, 12, 14, 17
3	2, 3, 5, 8
4	20, 21, 23, 26
5	29, 30, 32, 35
6	13, 15, 18
7	4, 6, 9
8	22, 24, 27
9	31, 33, 36
10	16, 19
11	7, 10
12	25, 28
13	34, 37

In segment $h=1$ no one in the household owns a licence, and so it is never possible for the household to acquire cars. In the other segments, shifts between *aext3* segments can occur when there are changes in total car ownership.

4.5 Model calibration

Section 4.2 set out the key formulae for the pivot approach, and introduced the 'sufficient statistics' technique as the calibration technique. This section details how the car ownership pivot models have been calibrated to match the Census target information.

4.5.1 Aggregate data

The aggregate data provides the aggregate choice proportions, namely T in Equation 4.3.

In order to get the run to work in ALOGIT, it is necessary to set up two dummy alternatives in the model, which have been termed 'PopCars' and 'PopNoCars'. These dummy alternatives are additional to the 0, 1, 2 and 3+ car alternatives. The utilities for these dummy alternatives are:

$$V(PopCars)_z = \beta_z \quad (4.5)$$

$$V(PopNoCars)_z = 0 \quad (4.6)$$

where: β_z is the zonal correction factor

The weight for the record is the total expanded QUAD household total for the zone. Thus the Census target data is specified as cars per household, which when weighted by the household total expands to the target number of cars.

The actual equations were slightly more complex, because the input file to the calibration procedure is a combined person-household file. Thus adjustments are required to take account of mean household size in order to achieve consistency with Equations 4.11 and 4.12, which are specified at the household level.

It is noted that because QUAD is an expansion, the expanded household total is not exactly equal to the target population. There is a discussion in Section 5.1 of these differences further for the 2006 base year.

4.5.2 Disaggregate data

In the sufficient statistics run, the disaggregate data corresponds to the right hand side of Equation 4.4. No choice information is specified from the disaggregate data.

For each household in the base sample in a given area type, it is necessary to define the probabilities P_{ni} of each car ownership state. These are defined by modifying the utilities of the *j* possible car ownership states as follows:

$$V(0cars)_z = \sum_s \beta_s X_s \quad (4.7)$$

$$V(1cars)_z = \sum_t \beta_t X_t + \beta_z \quad (4.8)$$

$$V(2cars)_z = \sum_u \beta_u X_u + 2\beta_z \quad (4.9)$$

$$V(3pcars)_z = \sum_v \beta_v X_v + c_3\beta_z \quad (4.10)$$

where: *z* is the model zone

$\beta_s X_s$ defines each of the *s* terms in the utility function for zero cars

$\beta_t X_t$ defines each of the *t* terms in the utility function for one car

$\beta_u X_u$ defines each of the *u* terms in the utility function for two cars

$\beta_v X_v$ defines each of the *v* terms in the utility function for three plus cars

β_z is the zonal correction factor

c_3 is the mean number of cars for the three-plus car alternative

The total car ownership models are documented in full in Tsang *et al* (2010). It is emphasised that the parameters from those models (β_s, β_t, β_u and β_v in Equations 4.7 to 4.10 above) are held constant during the calibration process. This means that for each zone only the calibration parameter β_z is estimated.

To define the weight associated with each record, it is necessary to combine information from the licence holding, company car ownership and QUAD model components. The weight is defined as:

$$w_n^{hp,cc} = \Pr_n^{hp} \Pr_n^{cc|hp} Q_n \quad (4.11)$$

where: $w_n^{hp,cc}$ is the weight for household *n* given *hp* and *cc*

Pr_n^{hp} is the probability of the household licence holding state *hp* (see Section 2.2.3)

$Pr_n^{cc|hp}$ is the probability of the company car ownership state *cc* (see Section 2.3.1) given *hp*

Q_n is the QUAD expansion factor for that household (see Section 2.5)

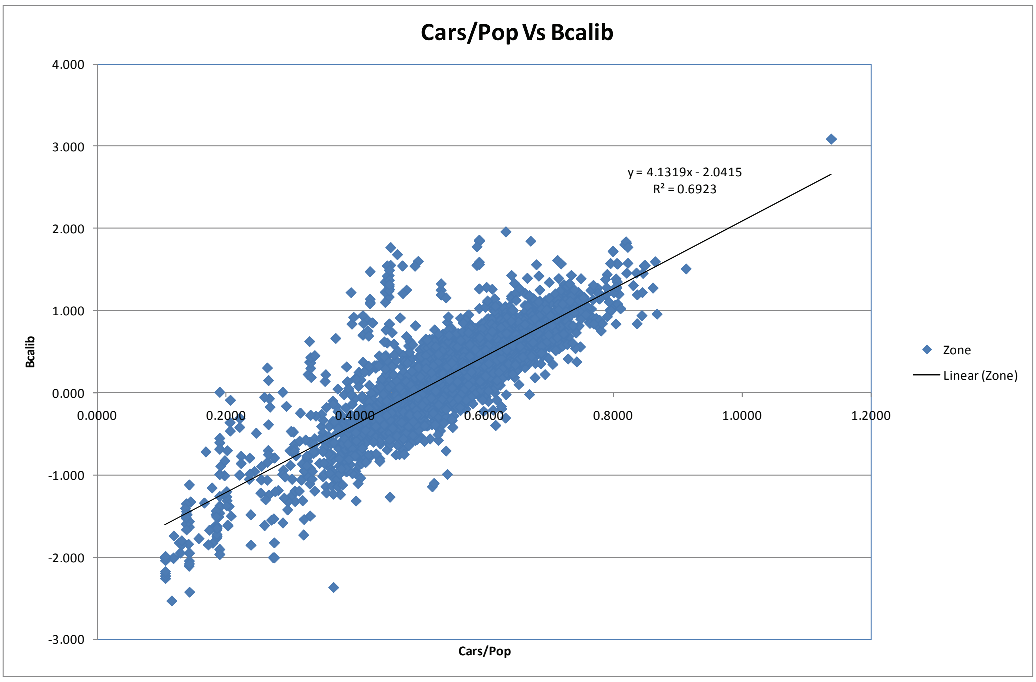
It is noted that for a given household *n*:

$$\sum_{hp}\sum_{cc} \Pr_n^{hp} \Pr_n^{cc|hp} = 1 \quad (4.12)$$

Thus the sum of $w_n^{hp.cc}$ over all households N is the total number of households in the zone predicted by QUAD.

A scatter plot of the zonal correction parameters obtained for the 2,690 model zones is presented in Figure 13. The x-axis is cars/population, rather than cars/adult, because of the way the information was extracted from the runs.

Figure 13: Zonal car ownership correction factor parameter scatter plot



The pattern in the zonal correction factors is what would be expected. For zones with low cars/population, negative correction factors are usually estimated to reduce the model predictions, whereas for zones with high cars/population positive correction factors are required to boost the model predictions up to the levels observed in the 2006 Census.

4.6 Model application

Equation 4.1 in Section 4.2 sets out the general form of the pivot-point equations, showing how changes in car ownership probabilities are predicted as incremental changes relative to the probabilities obtained prior to pivoting. This section details how Equation 4.1 has been implemented given the particular segmentations used in the STM.

The pivot has to be done separately for each combination of mode-destination segment (other than car ownership) and frequency segment and *therefore* separately by home-based purpose (because the segments are different). Of course it is also done separately by zone.

Suppose w_{kjz} is the weight in other-segment k and car ownership j, for zone z. Then the pivot is simply that:

$$w^*_{kjz} = \left(\sum_h w_{khz}\right) \frac{w_{kjz} \exp \beta_{jz}}{\sum_h w_{khz} \exp \beta_{hz}} \qquad (4.13)$$

where β_{jz} is the calibration value for ownership *j*, for zone *z*

k is a unique combination of mode-destination segment (other than car ownership) and frequency segment

Two points are necessary to show that this works correctly. First, the calibration value β_{jz} applies separately but equally to all of the segments – this is how the estimation is done and how it is applied.

Second, the total weight of the segment remains unchanged, because:

$$\sum_h w_{khz} = \sum_h w^*_{khz} \qquad (4.14)$$

$\sum_h w_{khz} = \sum_h w^*_{khz}$ which follows from the form of the Equation 4.13. Equation 4.14 simply states that the total weight before pivoting is equal to the total weight after pivoting.

Table 20 summarises the number of different car ownership pivot segments *K* by purpose. For a full definition of the mode destination and frequency segments, please refer to Chapter 3 of Fox *et al* (2011).

Table 20: Number of car ownership point segments *K* by purpose

Home-Based Purpose	Mode Destination Segments[12]	Frequency Segments	K
Work	10	18	180
Business	3	24	72
Primary education	2	4	8
Secondary education	1	2	2
Tertiary education	2	12	24
Shopping	6	36	216
Other travel	5	56	280

The car ownership model application has been validated by making special runs to compare the predicted car ownership post-pivoting to the Census target data for a number of selected test zones. The results of these tests for these selected test zones are summarised in Table 21.

12 Excluding car availability.

Table 21: Car ownership pivot application validation

STM Zone	Hhlds	Target Cars	Predicted Pre-pivot	Error	Predicted Post-pivot	Error
1289	752	775	998.2	28.8 %	774.8	0.0 %
1290	830	1,376	1,432.0	4.1 %	1,376.1	0.0 %
1291	115	198	190.3	-3.9 %	198.0	0.0 %
1292	33	49	52.7	7.6 %	49.0	0.0 %
1293	1,330	1,330	1,698.4	27.7 %	1,329.9	0.0 %
1294	1,292	1,686	1,931.9	14.6 %	1,686.2	0.0 %
1295	1,254	2,015	2,144.0	6.4 %	2,015.0	0.0 %
1297	11	5	12.5	150.1 %	4.8	-4.0 %
1298	176	225	249.7	11.0 %	225.0	0.0 %
1299	1,519	1,951	2,294.2	17.6 %	1,951.0	0.0 %
1300	1,447	2,801	2,760.1	-1.5 %	2,800.9	0.0 %

For zone 1297, which is a low population zone, a large adjustment is required, and so the model does not achieve an exact match to the target. For the other zones, where the numbers of cars are substantially higher, an excellent match to the target is achieved.

CHAPTER 5 Model runs and analysis

This chapter documents runs of the QUAD process that have been made for base and future years. Section 5.1 starts by setting out measures to allow the performance of the QUAD process to be assessed, and then documents a number of test runs that were made before deciding on the best base year QUAD run. Next, Section 5.2 describes the future year QUAD runs that have run for each of the forecast years (2011, 2016, 2021, 2026, 2031, 2036 and 2041). Finally, Section 5.3 documents analysis to validate the fit of QUAD at the spatial level using Geographical Information System (GIS) plots. This spatial validation has been undertaken for the 2006 base and 2036 forecast year QUAD runs.

5.1 Base year QUAD runs

A number of runs were made in order to determine the optimum expansion to the base year targets. Prior to documenting those runs in Section 5.1.2 onwards, Section 5.1.1 defines the performance measures that have been used to assess the base year expansion for each zone.

5.1.1 Performance measures

To define the performance measures, it is useful to restate Equation 2.1, which gives the QUAD objective function for each zone:

$$F(\varphi) = \sum_t w_t^a \left(y_t - \sum_c \varphi_c^a x_{tc}^a\right)^2 + \sum_c \left(\varphi_c^a - f_c^a\right)^2$$

It can be seen that the QUAD objective function comprises two elements, the first assessing the fit to the zonal targets, and the second the deviation from the base distribution. Root-mean square measures can be defined from these two elements as follows:

$$QF1^a = \sqrt{\frac{\sum_z \sum_t w_t^a \left(y_{tz} - \sum_c \varphi_{cz}^a x_{tc}^a\right)^2}{\sum_z \sum_t w_t^a}} \qquad (5.1)$$

$$QF2^a = \sqrt{\frac{\sum_z \sum_c \left(\varphi_{cz}^a - f_c^a\right)^2}{Z^a C}} \qquad (5.2)$$

$$TDEV^a = \sqrt{\sum_t \frac{1}{T}\left(\frac{\sum_z \left(y_{tz} - \sum_c \varphi_{cz}^a x_{ct}^a\right)}{\sum_z y_{tz}}\right)^2} \quad (5.3)$$

where: $QF1^a$ is the RMS measure across zones for the fit to the targets for area type a

$QF2^a$ is the RMS measure across zones for the fit to the household categories for area type a

$TDEV^a$ is an RMS measure of target deviation for area type a

w_t^a is the weight applied to target t

y_{tz} is the target value for zone z (from Census data)

φ_{cz} is the predicted frequency in the population for household category c in zone z

x_{ct}^a is the average quantity of target variable for a household of category c

f_c is the base frequency of household category c

T is the total number of targets

C is the total number of household categories

Z^a is the total number of zones in area type a

5.1.2 Student target runs

In the first set of runs all targets received a weight of 1. However, the error in new student target was 18.4%, compared with 6.3% across all targets. Therefore, a number of runs were made with the objective of improving the fit to the student target.

Table 22 summarises the results from these runs according to the fit measures set out in Section 5.1.1. The fit measures are calculated separately by area type, but weighted averages have been calculated for presentation in Table 22.

Table 22: Student target runs

Run No.	Run Description	Student Target Error	*TDEV*	*TDEV* (excl stud target)	QF1	QF2
11	all weights 1	18.4 %	6.3 %	5.1 %	0.023	0.010
12	student target weight 5, others weight 1	10.0 %	5.7 %	5.4 %	0.011	0.011
13	all weights 5	14.1 %	5.2 %	4.3 %	0.019	0.016
14	test run with student target weight 0, else 1	25.0 %	7.3 %	5.1 %	0.023	0.010

In run 12 the weight on the student target is increased to 5 to improve the fit for this target, and this results in a significant reduction in the error for the student target. The *TDEV* measure demonstrates the overall RMS measure improves with this change, though the second *TDEV* measure (which excludes the student target) shows some worsening for the other targets. By giving a higher weight to one of the targets, the fit to the base distribution would be expected to worsen, and indeed *QF2* shows a slight worsening relative to run 11.

In run 13, all targets receive a weight of 5, which ensures each target receives an equal weight. The student target fit is worse than run 12, but better than run 11. However, the

TDEV measures show an improvement relative to both runs 11 and 12. Further, the *QF2* measure demonstrates that the loss of fit relative to the base distribution is relatively modest.

Run 14 is a test run that demonstrates what happens if no attempt is made to match to the student target (a weight of zero). The *TDEV* measure shows the fit to the targets noticeably worse than the other runs, and as expected the error in fit to the student target worsens so that overall it is over-predicted by 25%.

Overall, it was concluded in discussion with the BTS that run 13 gave the best balance of fit to the targets and base distribution, and so run 13 has been taken as the agreed base year run.

5.1.3 Finalised base year run

Table 23 on the following page summarises the fit to the targets for run 13, the final base year run.

It can be seen that the age-gender targets are generally matched well, with errors of less than 5% in most cases. The largest errors occur for the 60+ targets: in inner Sydney for females (6.1% under-prediction) and in outer Sydney for males (7.1% over-prediction).

The household type targets show some variation between area types. In the middle and outer Sydney rings, couples with children and couples without children household types are predicted well, whereas in inner Sydney (where these household types are less common as a percentage of the total) these targets are matched less well. The 'other' household type, which is mixture of different household types, has the worst overall fit across all area types. It should be noted that this is the smallest of the five household types.

The full-time worker target is matched very well in inner and middle Sydney, and fairly well in outer Sydney and Newcastle & Wollongong. The part-time worker target is matched less well, but nonetheless the overall total is predicted within 2.6% of the target.

The first income target is a control total target to ensure children (individuals aged under 15) are predicted accurately, and this is indeed the case, with close predictions for each area type and an overall error of 0.1%. The lowest ($ 0–20,799) and middle ($ 31,200–41,599) income bands are predicted noticeably better than the other three bands, a pattern that is consistent across area types.

Table 23: Fit to zonal targets, run 13

Target		Inner Sydney			Middle Sydney Ring			Outer Sydney Ring			Newcastle, Wollongong			Total (summing over four area types)			Sydney		
No.	Name	Obs	Pred	% Error	Obs	Pred	% Error	Obs	Pred	% Error	Obs	Pred	% Error	Obs	Pred	% Error	Obs	Pred	% Error
1	M00_19	69,629	70,262	0.9%	147,250	147,188	0.0%	339,452	333,524	-1.7%	125,222	123,840	-1.1%	681,553	674,814	-1.0%	681,553	679,196	-0.3%
2	M20_39	161,557	154,106	-4.6%	172,713	170,009	-1.6%	318,970	328,049	2.8%	116,994	116,495	-0.4%	770,234	768,659	-0.2%	770,235	770,122	0.0%
3	M40_59	101,070	104,461	3.4%	147,074	146,612	-0.3%	308,383	309,537	0.4%	123,827	124,374	0.4%	680,354	684,984	0.7%	680,354	680,864	0.1%
4	M60+	57,029	58,338	2.3%	91,588	96,724	5.6%	164,143	175,789	7.1%	88,712	88,136	-0.6%	401,472	418,987	4.4%	401,471	418,750	4.3%
5	F00_19	67,027	67,824	1.2%	139,885	137,320	-1.8%	325,012	315,584	-2.9%	118,353	112,942	-4.6%	650,277	633,670	-2.6%	650,275	629,726	-3.2%
6	F20_39	160,965	160,206	-0.5%	173,032	175,655	1.5%	326,903	335,225	2.5%	114,871	119,252	3.8%	775,771	790,338	1.9%	775,770	793,786	2.3%
7	F40_59	99,313	100,480	1.2%	151,052	149,128	-1.3%	319,115	318,785	-0.1%	125,793	125,492	-0.2%	695,273	693,885	-0.2%	695,272	694,318	-0.1%
8	F60+	67,911	63,739	-6.1%	111,457	112,406	0.9%	194,866	190,941	-2.0%	103,719	104,286	0.5%	477,953	471,372	-1.4%	477,953	473,885	-0.9%
9	CplKid	76,895	79,031	2.8%	154,689	154,410	-0.2%	324,803	322,381	-0.7%	109,675	113,225	3.2%	666,062	669,047	0.4%	666,063	665,813	0.0%
10	CplOnl	92,827	85,785	-7.6%	97,505	97,564	0.1%	193,060	194,022	0.5%	98,969	97,057	-1.9%	482,361	474,428	-1.6%	482,362	479,731	-0.5%
11	1Parnt	27,290	30,118	10.4%	45,982	46,465	1.1%	104,580	111,222	6.4%	43,903	47,587	8.4%	221,755	235,392	6.1%	221,754	230,079	3.8%
12	Single	116,759	118,686	1.7%	92,262	94,446	2.4%	154,689	160,526	3.8%	87,174	87,555	0.4%	450,884	461,213	2.3%	450,884	461,837	2.4%
13	Oth HH	40,033	42,788	6.9%	22,639	24,814	9.6%	28,104	29,860	6.2%	14,702	15,037	2.3%	105,478	112,499	6.7%	105,477	112,716	6.9%
14	FT Workers	335,897	337,046	0.3%	384,309	384,827	0.1%	848,979	817,849	-3.7%	282,525	279,282	-1.1%	1,851,710	1,819,004	-1.8%	1,851,709	1,815,129	-2.0%
15	PT Workers	94,772	99,981	5.5%	150,116	143,049	-4.7%	301,873	317,406	5.1%	127,376	131,308	3.1%	674,137	691,744	2.6%	674,136	691,292	2.5%
16	Students	90,055	90,703	0.7%	103,774	115,852	11.6%	156,716	194,486	24.1%	62,460	70,341	12.6%	413,005	471,382	14.1%	413,005	468,174	13.4%
17	Inc_0(Age<15 years)	100,848	101,348	0.5%	214,235	216,747	1.2%	496,271	495,503	-0.2%	180,274	178,755	-0.8%	991,628	992,353	0.1%	991,627	996,702	0.5%
18	Inc_1 (<$20,800)	226,618	233,246	2.9%	419,874	418,018	-0.4%	781,467	784,864	0.4%	374,546	375,073	0.1%	1,802,505	1,811,201	0.5%	1,802,505	1,806,489	0.2%
19	Inc_2 ($20,800-$31,199)	72,651	65,577	-9.7%	120,917	115,131	-4.8%	258,893	226,853	-12.4%	107,254	93,350	-13.0%	559,715	500,911	-10.5%	559,715	488,877	-12.7%
20	Inc_3 ($31,200-$41,599)	69,778	70,208	0.6%	99,134	100,870	1.8%	214,917	211,316	-1.7%	75,404	72,957	-3.2%	459,233	455,351	-0.8%	459,232	449,913	-2.0%
21	Inc_4 ($41,600-$67,599)	140,725	154,992	10.1%	151,396	164,460	8.6%	325,034	371,598	14.3%	108,151	116,442	7.7%	725,306	807,492	11.3%	725,307	821,561	13.3%
22	Inc_5(>$67,599)	173,880	154,011	-11.4%	128,494	119,805	-6.8%	220,262	217,331	-1.3%	71,863	78,273	8.9%	594,499	569,420	-4.2%	594,499	576,991	-2.9%
			RMS	**5.5%**		**RMS**	**4.5%**		**RMS**	**7.3%**		**RMS**	**5.3%**		**RMS**	**5.2%**		**RMS**	**5.4%**

Analysis has been undertaken to compare the income distribution in the 2006 Census targets to the income distribution in the 2004-08 HTS data used for the base sample. This comparison is presented in Table 24.

Table 24: Comparison of Census and HTS personal income distributions

Income Band	2006 Census		2004-2008 HTS		Diff
< $20,800	1,802,505	43.5%	1,758,359	42.4%	-1.1%
$20,800-31,199	559,715	13.5%	482,622	11.6%	-1.9%
$31,200-41,599	459,232	11.1%	450,654	10.9%	-0.2%
$41,600-67,599	725,307	17.5%	847,531	20.4%	2.9%
> $67,599	594,499	14.4%	607,088	14.6%	0.3%
Total	4,141,258	100.0%	4,146,254	100.0%	0.0%

It can be seen from Table 24 that the Census and HTS data distributions match well overall, though the percentage of persons in the fourth income band ($ 41,600–67,599) is higher in the HTS distribution than in the Census. This difference explains why the income target is matched less well for this particular income band, i.e. there are more people in the HTS distribution in this band, and this results in an over-prediction for this band relative to the target value from the Census.

5.1.4 Impact of area type

For run 13, a test run was made using a single set of inputs for the whole study area, rather than running four separate runs by area type as set out in Section 2.5. The *TDEV* measure was 5.4%, compared with 5.2% when the results from separate runs by area type were summed. Therefore running QUAD separately by area type does give an improved fit to the targets, but the improvement is relatively modest.

However, using the *QF2* measure, which measures the deviation from the base sample(s), larger benefits were observed. This measure was 40.9 summing across the four area types in run 13, whereas in the all study area run the measure was 50.3, i.e. 23% worse. Thus, as would be expected, a closer match to the base distributions is achieved by running separately by the four area types.

5.2 Future year QUAD runs

Future year runs have been made using zonal target data supplied by the BTS for 2011, 2016, 2021, 2026, 2031, 2036 and 2041. Following on from the test runs described in Section 5.1.2 all the target weights in the future runs receive a weight of five, and following the test runs described in Section 5.1.4 separate runs are made by area type.

5.2.1 Income growth

An important consideration when running QUAD for future years is to determine the appropriate welfare factor, which gives the balance between income growth resulting from increases in general welfare, and growth due to re-distribution between income bands. The thinking behind the welfare factor approach is discussed in more detail in Section 3.3.2.

The strategy adopted for this work was to start by making a series of runs for 2036, which represents a 30 year forecasting period from the base, and to examine the impact of different assumptions for the welfare increase.

The assumption of 1% growth in real income per annum implies a 34.8% growth in incomes in real terms between 2006 and 2036. A number of QUAD runs were undertaken using different assumptions for the welfare factor to best match this 34.8% income growth. As noted in Section 3.2.3, the assumption of 1% real terms growth in incomes per annum used to derive the income targets did not take into account demographic effects, and in particular population ageing, which will reduce the proportion of the population of working age and hence average incomes. If these effects were to be included, the total income increase between 2006 and 2036 would be expected to be somewhat lower than 34.8%.

These runs are summarised in Table 25, which shows the overall fit to the targets and the base sample using the measures defined in Section 5.1.1, and Table 26, which shows the percentage error in the fit to the income and worker targets. The worker targets are presented because our experience is that the welfare factor has a significant impact on the fit to worker targets because a household's income is strongly correlated with the number of workers in the household.

Table 25: 2036 income growth QUAD runs, overall fit

Welfare factor	*QF1* Target deviation	*QF2* Base sample dev.	*TDEV* Target total deviation
1.348	0.0195	0.0170	3.7 %
1.300	0.0193	0.0170	2.7 %
1.250	0.0219	0.0181	9.5 %

Table 26: 2036 income growth runs, income and worker target fit

Welfare factor	Income targets ($k)					Worker targets	
	< 20.8	20.8–31.2	31.2–41.6	41.6–67.6	> 67.6	Full time	Other
1.348	0.7 %	-6.4 %	10.7 %	-0.5 %	-5.3 %	2.5 %	-4.0 %
1.300	0.6 %	3.2 %	-1.1 %	-0.3 %	-5.3 %	2.7 %	-3.5 %
1.250	5.8 %	-36.4 %	-12.8 %	1.4 %	-5.5 %	4.5 %	-0.4 %

From Table 25 it can be seen that a welfare factor of 1.3 minimises both the deviation from the targets (*QF1* and *TDEV*) and the deviation from the base sample (*QF2*).

Table 26 demonstrates that the fit to the income targets is best for a welfare factor of 1.3. The full-time worker target is slightly better predicted with a welfare factor of 1.348 (i.e. no re-distribution) but the loss of fit to the worker targets in moving from 1.348 to 1.3 is modest. For a welfare factor of 1.25, the fit to the full-time worker target worsens noticeably. On the basis of overall fit to the data, and fit to the income and worker targets, a welfare factor of 1.30 was selected for 2036.

The welfare factors used for each of the forecast years were determined using a similar approach. Table 27 summarises the welfare factor for each year, and shows the balance between the welfare increase and income re-distribution.

Table 27: Welfare factors by year

Year	Total Income Increase	Welfare Factor	Redistribution	Percentage Redistribution
2006	1.000	1.000	0.000	0.0 %
2011	1.051	1.048	0.003	0.3 %
2016	1.105	1.097	0.008	0.7 %
2021	1.161	1.146	0.015	1.3 %
2026	1.220	1.197	0.023	1.9 %
2031	1.282	1.250	0.033	2.6 %
2036	1.348	1.300	0.048	3.5 %
2041	1.417	1.347	0.069	4.9 %

It can be seen that the majority of the income increase is explained by the general welfare increase, with only modest levels of redistribution between income bands. The proportion of the income increase explained by redistribution increases steadily with time.

5.2.2 Fit to targets

To assess the overall fit of the future year QUAD runs, the *QF1*, *QF2* and *TDEV* measures defined in Section 5.1.1 have been used. The fit of the future year runs assessed using these measures is presented in Table 28. It is noted that the run for 2036 is later than the runs documented in Section 5.2.1 (which were obtained prior to the use of imputed incomes, described in Section 3.3.3) and therefore different fit measures are presented.

Table 28: Future year QUAD runs, overall fit

Year	*QF1* Target deviation	*QF2* Base sample dev.	*TDEV* Target total deviation
2006	0.0191	0.0160	5.18 %
2011	0.0188	0.0159	4.25 %
2016	0.0185	0.0158	3.50 %
2021	0.0182	0.0161	3.05 %
2026	0.0185	0.0165	3.35 %
2031	0.0192	0.0168	4.13 %
2036	0.0197	0.0171	4.62 %
2041	0.0201	0.0172	5.03 %

It can be seen that the fit to the targets is actually slightly better than for the 2006 base for the 2011 to 2026 forecast years, and then for years after 2031 deteriorates slightly. The fit to the base sample measure shows little change up to 2021, and then some deterioration for 2026 and onwards, which would be expected for forecast years further from the base. The total target deviation measure shows better fit than the 2006 base for all forecast years, with the best fit to the targets observed for 2021. In summary, measures of overall fit are comparable (or better) than those obtained for the base year, and so were judged acceptable.

The fit to the income targets and worker targets was also examined by calculating the percentage difference between the predicted and target totals for each income band. These results are presented for base and forecast years in Table 29.

Table 29: Fit to income and worker targets for base and forecast years

Welfare Factor	Income Targets ($k)					Worker Targets	
	< 20.8	20.8–31.2	31.2–41.6	41.6–67.6	> 67.6	Full time	Other
2006	0.5 %	-10.5 %	-0.8 %	11.3 %	0.5 %	-1.7 %	2.6 %
2011	0.5 %	-6.9 %	-3.5 %	7.3 %	0.5 %	-1.2 %	2.6 %
2016	0.6 %	-3.6 %	-4.7 %	4.1 %	0.6 %	0.5 %	1.5 %
2021	0.5 %	0.6 %	-4.6 %	2.2 %	0.5 %	1.5 %	-1.2 %
2026	0.6 %	4.4 %	-8.5 %	0.5 %	0.6 %	1.1 %	-1.8 %
2031	0.5 %	9.5 %	-11.6 %	-2.1 %	0.5 %	2.8 %	-1.3 %
2036	0.5 %	12.6 %	-12.3 %	-2.7 %	0.5 %	2.7 %	-3.5 %
2041	0.0 %	15.8 %	-11.9 %	-4.0 %	0.0 %	3.5 %	-3.5 %

For the lowest income band the predicted totals match the targets closely for all years. For the $ 20,800–31,199 band, the target is under-predicted in 2006, then this under-prediction declines until the target is predicted accurately in 2021. For subsequent years, this band is over-predicted. The middle income band is predicted well in 2006, but then under-predicted in 2011 and beyond. The $ 41,600–67,599 band is over-predicted in 2006, but better predicted for subsequent years. Finally, the top income band is predicted reasonably well for all forecast years.

The fit to the worker targets is fairly good for all years. The best fit is obtained for 2016, and then for later forecast years there is a pattern of increasing over-prediction of full-time workers, and under-prediction of other worker types.

In summary, the top and bottom income bands are forecast well for all years. For the middle three income bands, some differences exist between target and predicted values, but the fit to income bands is judged to be reasonable. The fit to the worker targets is reasonably good for all years. Noting that the overall fit to the targets is better than the base case using the *TDEV* measure, it was decided to accept these future target forecasts.

5.3 Detailed spatial validation

The QUAD validation presented in Sections 5.1 and 5.2 assessed the fit for base and future year QUAD runs by comparing target and predicted values summed across the whole study area. To validate the ability of the QUAD process to match the target variables at the spatial level, GIS plots have been generated to compare target and predicted information at the zonal level.

This spatial validation has been undertaken for both the 2006 base and 2036 forecast years, and for each of the 22 target variables. Furthermore, for each combination of year and target variable, two plots have been produced, one covering the entire study area, and another focused on Sydney to show the differences for Sydney in more detail. This means that a total of 88 plots have been generated. Given the volume of output, these plots are presented in a separate appendix volume, rather than in this report. This section presents analysis undertaken to generate summary measures of spatial fit for each target variable, and then discusses patterns of difference in the GIS plots for those targets where the fit to the target data is less good.

When the GIS plots were initially generated, the percentage differences between target and predicted values were plotted. However, the impact of large percentage differences for a given target was hard to assess, because large differences may relate to zones with low population totals, where differences are of less concern relative to well populated zones. Therefore it was decided to use a GEH measure, more typically employed in comparing link flows to validate highway assignments. The GEH measure can be defined as (DMRB, 1996):

$$GEH = \sqrt{\frac{(P-T)^2}{0.5*(P+T)}}$$

where: *P* is the predicted amount of the variable

T is the target amount of the variable

The GEH measure has the advantage that it accounts for scale, so in the context of the QUAD validation, differences for zones with higher amounts of the target variable have proportionally more impact.

The criterion used in DMRB in the context of highway assignment is that 85% of links should have a GEH of less than 5. An analogous criterion has been applied to the spatial QUAD validation, with the condition tested that for a given target 85% of zones should have a GEH of 5 or less.

5.3.1 2006 base year

Table 30 presents the results of the GEH analysis of fit for each of the target variables.

Table 30: QUAD GEH validation, 2006 base

Target	% Zones with GEH ≤ 5	% Zones with GEH > 5
Males, 0–19	99.8 %	0.2 %
Males, 20–39	99.6 %	0.4 %
Males, 40–59	99.9 %	0.1 %
Males, 60+	99.2 %	0.8 %
Females, 0–19	99.8 %	0.2 %
Females, 20–39	99.8 %	0.2 %
Females, 40–59	99.8 %	0.2 %
Females, 60+	98.7 %	1.3 %
Couples with children households	99.6 %	0.4 %
Couple only households	99.7 %	0.3 %
Single parent households	99.0 %	1.0 %
Single person households	99.8 %	0.2 %
Other household types	89.7 %	10.3 %
Full-time workers	99.7 %	0.3 %
Part-time workers	98.3 %	1.7 %
Students	91.5 %	8.5 %
Aged < 15	99.9 %	0.1 %
Income < $20,800 p.a.	99.9 %	0.1 %
Income $20,800–31,199 p.a.	96.7 %	3.3 %
Income $31,200–41,599 p.a.	99.3 %	0.7 %
Income $41,600–67,599 p.a.	96.2 %	3.8 %
Income > $67,599 p.a.	68.0 %	32.0 %

Most of the 22 target variables pass the GEH test comfortably. There are three targets where the GEH measure indicates significant differences between target and predicted values at the spatial level, the other household types target and the new students target (both of which do not actually fail the GEH test but do perform noticeably worse than the other targets), and the top income band which fails the test.

For the other households type target, the GIS plots show no clear pattern in the particular location of high GEH zones across Sydney. There is a tendency for Inner Sydney to have more zones with higher GEH values. This percentage of households of this type is higher in Inner Sydney, and so the absolute numbers of households are higher, which will result in higher GEH measures.

For the students target, the GIS plots show differences for both inner and more suburban zones, and no clear patterns emerge where differences are concentrated in particular areas.

For the top income target, the GIS plots show that the high difference zones are concentrated in Inner Sydney, and in particular to areas to the north of the Central Business District (CBD) such as Chatswood and Mosman. Zones in these areas will contain a high fraction of high income households, and it is probable that QUAD is under-predicting the target totals for these zones, resulting in the high GEH values.

5.3.2 2036 forecast year

Table 31 presents the results of the GEH analysis of fit for each of the target variables.

Table 31: QUAD GEH validation, 2036 base

Target	% Zones with GEH ≤ 5	% Zones with GEH > 5
Males, 0–19	99.7 %	0.3 %
Males, 20–39	98.7 %	1.3 %
Males, 40–59	99.4 %	0.6 %
Males, 60+	96.4 %	3.6 %
Females, 0–19	99.5 %	0.5 %
Females, 20–39	99.0 %	1.0 %
Females, 40–59	99.7 %	0.3 %
Females, 60+	96.8 %	3.2 %
Couple with Children Households	97.5 %	2.5 %
Couple only households	99.0 %	1.0 %
Single parent households	98.1 %	1.9 %
Single person households	99.6 %	0.4 %
Other household types	86.2 %	13.8 %
Full-time workers	94.9 %	5.1 %
Part-time workers	98.8 %	1.2 %
Students	90.8 %	9.2 %
Aged < 15	100.0 %	0.0 %
Income < $20,800 p.a.	99.0 %	1.0 %
Income $20,800–31,199 p.a.	93.2 %	6.8 %
Income $31,200–41,599 p.a.	90.2 %	9.8 %
Income $41,600–67,599 p.a.	97.5 %	2.5 %
Income > $67,599 p.a.	72.2 %	27.8 %

The results for the 2036 forecast year are similar to the 2006 results, with most targets comfortably passing the GEH test. The other household types target has a fit that is noticeably worse than the other targets, and comes close to failing the GEH test, and the fit of the students target is also poor relative to the other targets. The fit for the middle income band is relatively poor, and once again the top income band target is the worst fitting target by far.

For the other household type target, higher GEH zones are distributed across the study area, with no obvious concentrations in particular areas.

For the students target, the story is similar, with higher GEH zones spread across the study area and no particularly high difference areas standing out.

For the middle income band, larger GEH values are found in middle and outer Sydney, and in Newcastle and Wollongong, with Inner Sydney predicted relatively well.

Finally, for the top income band, as per the 2006 validation, the highest differences are again observed in the area to the north of the CBD covering Chatswood and Mosman.

CHAPTER 6 Summary

Updates to the synthesiser

The Population Synthesiser has been updated to reflect a 2006 base year. The new prototypical sample is drawn from 2004-08 Household Travel Survey data, and contains 24,475 persons from 9,915 households.

Base year target data for the synthesiser has been assembled from 2006 Census data. A special procedure has been used for low population zones whereby households are transferred in from neighbouring zones to ensure that each zone contains at least 10 households.

The new 2006 base licence and car ownership models have been implemented in the new Population Synthesiser. To implement these models, the implementations have been revised to read in the new prototypical sample data, read in the new model parameters, take account of changes to the model specifications, and make adjustments so that costs are specified in 2006 prices.

A special car availability procedure is used in model application so that changes in accessibility between future and base years have an impact on the car ownership levels predicted by the Population Synthesiser. This procedure has been re-calibrated for the new 2006 base version of the STM.

A procedure named 'QUAD' is used to expand the Prototypical Sample to best meet target variables defined for each model zone. A number of improvements have been made to the procedure used to apply QUAD. First, the treatment of incomes has been improved so that income growth is now decomposed into a general welfare increase that increases all incomes by a uniform amount and re-distribution between income bands. Second, the procedure is now run separately by four area types, which takes account of differences in the characteristics of the population across the study area. As a result of splitting the prototypical sample into the four area types, the number of household categories used in the QUAD expansion has been reduced from 74 to 59. Third, the target definitions have been updated, with a new student target, manufacturing and non-manufacturing targets replaced by full and part-time workers, and the number of personal income targets increased from four to five.

The final stage in the Population Synthesiser is to generate files that define the population by zone, segment and purpose, ready for application in the Travel Demand models. The procedure that performs this step, 'ACCUM', has been updated to reflect the segmentations used in the 2006 base versions of the Travel Demand models.

Car ownership pivot

A significant improvement introduced in the 2006 base Population Synthesiser is a car ownership pivoting step.

Validation work undertaken for the 2001 base Population Synthesiser demonstrated that the car ownership model was not able to predict the full range in car ownership levels observed across zones. Therefore the new car ownership pivot process was developed to pivot off the car ownership levels by zone observed in the 2006 Census.

To implement the car ownership pivot, a correction factor is calculated for each zone, which ensures that when the Population Synthesiser is applied for the 2006 base year car ownership levels match those observed in the 2006 Census exactly. When the model is applied for future years the same correction factors are incorporated into the model predictions.

QUAD validation

As noted above, QUAD is the process used to expand the prototypical sample to best meet target variables defined for each model zone.

The target variables cover age-gender bands, household types, worker types, students and income bands. The ability of QUAD to match these targets is key to the ability of the Population Synthesiser to predict accurately the number of persons by zone and segment. Therefore a number of checks have been undertaken to validate the performance of the QUAD process.

The first set of runs was made for the 2006 base year. For the new student target, the match at the zonal level between target and predicted values was poor relative to the other target variables. Following a number of different tests, the weights for *all* of the targets were increased from 1 to 5. This resulted in an improved fit to the student target, as well as an improved fit to the other targets, for an acceptable loss of fit to the base sample distribution. Therefore weights of 5 have been retained as the standard for the 2006 base Population Synthesiser.

A test was made to assess the impact of running the QUAD procedure separately by the four area types. An improved fit to the zonal targets was observed, though the difference was relatively modest. Examining the fit to the target variables at the area type level flagged that the other household types target, the new students target noted above, and some of the income band targets fit less well than the other target variables.

Future year QUAD runs have been made for 2011, 2016, 2021, 2026, 2031, 2036 and 2041. To make these runs, appropriate 'welfare factors' were required, which give the balance between the welfare increase applied to all incomes, and re-distribution between bands. For 2036, the total real terms growth in income is 34.8%. Test runs demonstrated that the best fit to the targets and base sample was achieved by assuming a 30% welfare increase, and just 3.8% redistribution between income bands. For the other forecast years, the welfare factor has also been found to be the dominant effect.

To validate the future year runs, the overall fit to the five income and two worker targets was reviewed. The fit varied somewhat with year, but overall the middle three income bands fit less well than the top and bottom income bands and the two worker targets.

The final stage in the QUAD validation was to use GIS plots to compare target and predicted values at the zonal level, with GEH measures used to provide a measure of difference that takes account of differences in scale in the target variables between model zone. Plots were produced for each of the 22 zonal targets for the 2006 base and 2036 forecast years, and given the number of plots generated these are presented in a separate volume of appendices.

To summarise the GIS analysis, the percentage of zones achieving a certain GEH value was calculated for each of the zonal targets. Most targets comfortably passed the test criterion that 85% of zones should have a GEH of 5 of lower. The targets with a poorer spatial fit for the 2006 base year were the other household type target, the new students target, and the top income band, where differences were noted in the area immediately to the north west of the city centre. The pattern of differences for the 2036 forecast year was similar, but in addition for the middle income band significant differences were observed in middle and outer Sydney, and in Newcastle and Wollongong.

REFERENCES

Reference list

Daly, A. (1982) Estimating Choice Models Containing Attraction Variables, *Transportation Research*, Vol. 16B, 1.

DMRB (1996) *Design Manual for Roads and Bridges, Part 1: Traffic Appraisal in Urban Areas*, Department of Transport, London.

Fox, J., A. Daly and M. Kouwenhoven (2004) *Sydney Strategic Travel Model: Stage 2 Validation, Population Model*, Report RED-03136-1, RAND Europe, Cambridge, UK.

Fox, J., A. Daly and B. Patruni (2010) *Sydney Strategic Model Re-estimation: Mode-Destination Model*, Report DRR-5270-BTS, RAND Europe, Cambridge, UK.

Fox, J., B. Patruni and A. Daly (2011) *Application System for Sydney Strategic Model*, Report TR-949-BTS, RAND Europe, Cambridge, UK.

HCG and ITS (2000) *Implementation of the Work Model*, Report 9009-4, Hague Consulting Group, Cambridge, UK.

HCG and ITS (2002) *Model Implementation*, Report 0032-7, Hague Consulting Group, Cambridge, UK.

Raimond, T. and F. Milthorpe (2010) *Why Are Young People Driving Less? Trends in Licence-holding and Travel Behaviour*, paper presented at 2010 Australasian Transport Research Forum, Canberra, Australia.

TDC (2009) *Household Travel Survey (HTS) Database Documentation*, Transport Data Centre, Transport and Infrastructure, NSW Government.

Train, K. (2003) *Discrete Choice Models with Simulation*, Cambridge University Press, UK.

Tsang, F., A. Daly, J. Fox and B. Patruni (2010) *Sydney Strategic Model Re-estimation: Licence, Car Ownership and Frequency Models*, Report DRR-5271-BTS, RAND Europe, Cambridge, UK.

APPENDICES

Appendix A: Merging of household categories

Table A1: Cross-tabulation of the original 74 household categories CAT by area

	Area				
CAT	Inner	Middle	Outer	Regional	Total
1	6	5	7	4	22
2	61	34	47	33	175
3	2	6	76	52	136
4	2	15	37	14	68
5	25	9	15	20	69
6	157	87	107	59	410
7	21	78	118	57	274
8	67	79	208	99	453
9	44	60	78	54	236
10	154	122	181	99	556
11	7	13	26	14	60
12	44	31	63	33	171
13	123	208	265	160	756
14	31	29	39	31	130
15	10	18	14	4	46
16	56	50	40	20	166
17	210	184	204	118	716
18	3	7	29	13	52
19	9	51	152	73	285
20	9	51	135	41	236
21	6	6	6	8	26
22	62	72	72	24	230
23	378	226	268	76	948
24	0	83	113	29	225
25	270	659	1,018	319	2,266
26	595	929	1,904	747	4,175
27	58	84	162	84	388
28	112	132	304	112	660
29	240	252	502	260	1,254
30	9	28	41	0	78
31	103	151	180	50	484
32	271	364	604	259	1,498
33	156	360	515	280	1,311
34	58	88	136	68	350
35	32	64	81	36	213
36	27	33	30	9	99
37	30	51	48	24	153
38	75	57	105	36	273
39	0	13	21	0	34
40	16	12	24	5	57

	Area				
CAT	Inner	Middle	Outer	Regional	Total
41	4	25	37	13	79
42	15	33	57	21	126
43	27	21	30	9	87
44	27	45	45	15	132
45	19	42	50	24	135
46	27	80	100	22	229
47	27	43	136	65	271
48	24	45	78	33	180
49	24	108	129	63	324
50	93	114	222	81	510
51	22	53	81	12	168
52	61	87	140	18	306
53	48	138	257	133	576
54	12	31	54	26	123
55	10	9	12	13	44
56	13	3	22	5	43
57	8	45	29	17	99
58	31	51	34	8	124
59	26	38	97	16	177
60	15	25	46	0	86
61	0	13	28	6	47
62	6	5	95	13	119
63	8	29	8	9	54
64	0	12	25	12	49
65	19	18	41	17	95
66	18	77	52	24	171
67	0	20	17	11	48
68	18	11	42	13	84
69	8	63	42	20	133
70	20	70	104	20	214
71	52	124	254	70	500
72	28	82	39	17	166
73	28	44	87	10	169
74	33	79	164	62	338
				Total	25,745

Table A2: Cross-tabulation of the new 59 household categories CAT2 by area

	Area				
CAT2	Inner	Middle	Outer	Regional	Total
1	8	11	83	56	158
2	63	49	84	47	243
3	25	9	15	20	69
4	157	87	107	59	410
5	21	78	118	57	274
6	67	79	208	99	453
7	44	60	78	54	236
8	154	122	181	99	556
9	7	13	26	14	60
10	44	31	63	33	171
11	123	208	265	160	756
12	31	29	39	31	130
13	13	25	43	17	98
14	56	50	40	20	166
15	210	184	204	118	716
16	9	51	152	73	285
17	9	51	135	41	236
18	6	89	119	37	251
19	62	72	72	24	230
20	378	226	268	76	948
21	270	659	1018	319	2,266
22	595	929	1904	747	4,175
23	67	112	203	84	466
24	112	132	304	112	660
25	240	252	502	260	1,254
26	103	151	180	50	484
27	271	364	604	259	1,498
28	156	360	515	280	1,311
29	58	88	136	68	350
30	32	64	81	36	213
31	27	33	30	9	99
32	30	51	48	24	153
33	75	57	105	36	273
34	20	50	82	18	170
35	15	33	57	21	126
36	27	21	30	9	87
37	27	45	45	15	132
38	19	42	50	24	135
39	27	80	100	22	229
40	27	43	136	65	271
41	24	45	78	33	180
42	24	108	129	63	324
43	93	114	222	81	510
44	22	53	81	12	168
45	61	87	140	18	306
46	48	138	257	133	576
47	35	43	88	44	210
48	8	45	29	17	99
49	57	89	131	24	301
50	21	43	169	19	252
51	27	59	74	38	198
52	18	77	52	24	171
53	18	31	59	24	132
54	8	63	42	20	133

	Area				
CAT2	Inner	Middle	Outer	Regional	Total
55	20	70	104	20	214
56	52	124	254	70	500
57	28	82	39	17	166
58	28	44	87	10	169
59	33	79	164	62	338
				Total	25,745

Appendix B: Area type definitions

The table presented overleaf provides a full definition of the four area types used in the model.

Ring No	Ring Name	SD2006 No	SD2006 Name	SSD2006 No	SSD2006 Name	SLA2006 No	SLA2006 Name	No.of Zones
1	Inner	5	Sydney	505	Inner Sydney	1100	Botany Bay	23
1	Inner	5	Sydney	505	Inner Sydney	4800	Leichhardt	27
1	Inner	5	Sydney	505	Inner Sydney	5200	Marrickville	37
1	Inner	5	Sydney	505	Inner Sydney	7201	Sydney - Inner	166
1	Inner	5	Sydney	505	Inner Sydney	7204	Sydney - East	36
1	Inner	5	Sydney	505	Inner Sydney	7205	Sydney - South	42
1	Inner	5	Sydney	505	Inner Sydney	7206	Sydney - West	19
1	Inner	5	Sydney	510	Eastern Suburbs	6550	Randwick	52
1	Inner	5	Sydney	510	Eastern Suburbs	8050	Waverley	23
1	Inner	5	Sydney	510	Eastern Suburbs	8500	Woollahra	27
1	Inner	5	Sydney	535	Inner Western Sydney	150	Ashfield	17
1	Inner	5	Sydney	535	Inner Western Sydney	1521	Canada Bay - Concord	19
1	Inner	5	Sydney	535	Inner Western Sydney	1524	Canada Bay - Drummoyne	13
1	Inner	5	Sydney	555	Lower Northern Sydney	4700	Lane Cove	16
1	Inner	5	Sydney	555	Lower Northern Sydney	5350	Mosman	13
1	Inner	5	Sydney	555	Lower Northern Sydney	5950	North Sydney	36
2	Middle	5	Sydney	515	St George-Sutherland	4150	Hurstville	31
2	Middle	5	Sydney	515	St George-Sutherland	4450	Kogarah	24
2	Middle	5	Sydney	515	St George-Sutherland	6650	Rockdale	42
2	Middle	5	Sydney	520	Canterbury-Bankstown	351	Bankstown - North-East	23
2	Middle	5	Sydney	520	Canterbury-Bankstown	353	Bankstown - North-West	32
2	Middle	5	Sydney	520	Canterbury-Bankstown	355	Bankstown - South	26
2	Middle	5	Sydney	520	Canterbury-Bankstown	1550	Canterbury	47
2	Middle	5	Sydney	535	Inner Western Sydney	1300	Burwood	16
2	Middle	5	Sydney	535	Inner Western Sydney	7100	Strathfield	21
2	Middle	5	Sydney	540	Central Western Sydney	200	Auburn	35
2	Middle	5	Sydney	540	Central Western Sydney	6251	Parramatta - Inner	34
2	Middle	5	Sydney	540	Central Western Sydney	6252	Parramatta - North-East	23
2	Middle	5	Sydney	540	Central Western Sydney	6253	Parramatta - North-West	17
2	Middle	5	Sydney	540	Central Western Sydney	6254	Parramatta - South	16
2	Middle	5	Sydney	555	Lower Northern Sydney	4100	Hunter's Hill	8
2	Middle	5	Sydney	555	Lower Northern Sydney	6700	Ryde	64
2	Middle	5	Sydney	555	Lower Northern Sydney	8250	Willoughby	32
2	Middle	5	Sydney	560	Central Northern Sydney	4500	Ku-ring-gai	43
2	Middle	5	Sydney	565	Northern Beaches	5150	Manly	15
3	Outer	5	Sydney	515	St George-Sutherland	7151	Sutherland Shire - East	42
3	Outer	5	Sydney	515	St George-Sutherland	7152	Sutherland Shire - West	39
3	Outer	5	Sydney	525	Fairfield-Liverpool	2851	Fairfield - East	48
3	Outer	5	Sydney	525	Fairfield-Liverpool	2854	Fairfield - West	34
3	Outer	5	Sydney	525	Fairfield-Liverpool	4901	Liverpool - East	47
3	Outer	5	Sydney	525	Fairfield-Liverpool	4904	Liverpool - West	61
3	Outer	5	Sydney	530	Outer South Western Sydney	1450	Camden	66
3	Outer	5	Sydney	530	Outer South Western Sydney	1501	Campbelltown - North	42
3	Outer	5	Sydney	530	Outer South Western Sydney	1504	Campbelltown - South	38
3	Outer	5	Sydney	530	Outer South Western Sydney	8400	Wollondilly	17
3	Outer	5	Sydney	540	Central Western Sydney	3950	Holroyd	41
3	Outer	5	Sydney	545	Outer Western Sydney	900	Blue Mountains	37
3	Outer	5	Sydney	545	Outer Western Sydney	3800	Hawkesbury	33
3	Outer	5	Sydney	545	Outer Western Sydney	6351	Penrith - East	48
3	Outer	5	Sydney	545	Outer Western Sydney	6354	Penrith - West	41
3	Outer	5	Sydney	553	Blacktown	751	Blacktown - North	62
3	Outer	5	Sydney	553	Blacktown	752	Blacktown - South-East	45
3	Outer	5	Sydney	553	Blacktown	753	Blacktown - South-West	42
3	Outer	5	Sydney	560	Central Northern Sydney	501	Baulkham Hills - Central	38
3	Outer	5	Sydney	560	Central Northern Sydney	503	Baulkham Hills - North	39
3	Outer	5	Sydney	560	Central Northern Sydney	505	Baulkham Hills - South	17
3	Outer	5	Sydney	560	Central Northern Sydney	4001	Hornsby - North	29
3	Outer	5	Sydney	560	Central Northern Sydney	4004	Hornsby - South	41
3	Outer	5	Sydney	565	Northern Beaches	6370	Pittwater	29
3	Outer	5	Sydney	565	Northern Beaches	8000	Warringah	51
3	Outer	5	Sydney	570	Gosford-Wyong	3101	Gosford - East	28
3	Outer	5	Sydney	570	Gosford-Wyong	3104	Gosford - West	45
3	Outer	5	Sydney	570	Gosford-Wyong	8551	Wyong - North-East	30
3	Outer	5	Sydney	570	Gosford-Wyong	8554	Wyong - South and West	32
4	Regional	10	Hunter	1005	Newcastle	1720	Cessnock	27
4	Regional	10	Hunter	1005	Newcastle	4651	Lake Macquarie - East	21
4	Regional	10	Hunter	1005	Newcastle	4653	Lake Macquarie - North	33
4	Regional	10	Hunter	1005	Newcastle	4655	Lake Macquarie - West	24
4	Regional	10	Hunter	1005	Newcastle	5050	Maitland	24
4	Regional	10	Hunter	1005	Newcastle	5903	Newcastle - Inner City	23
4	Regional	10	Hunter	1005	Newcastle	5904	Newcastle - Outer West	16
4	Regional	10	Hunter	1005	Newcastle	5905	Newcastle - Throsby	23
4	Regional	10	Hunter	1005	Newcastle	6400	Port Stephens	24
4	Regional	15	Illawarra	1505	Wollongong	4400	Kiama	9
4	Regional	15	Illawarra	1505	Wollongong	6900	Shellharbour	27
4	Regional	15	Illawarra	1505	Wollongong	8451	Wollongong - Inner	43
4	Regional	15	Illawarra	1505	Wollongong	8454	Wollongong Bal	53
4	Regional	15	Illawarra	1507	Nowra-Bomaderry	6951	Shoalhaven - Pt A	16
4	Regional	15	Illawarra	1510	Illawarra SD Bal	6952	Shoalhaven - Pt B	27
4	Regional	15	Illawarra	1510	Illawarra SD Bal	8350	Wingecarribee	23

Appendix C: Technical manual

Running the Population Synthesiser

The Population Synthesiser is run from a batch file named 'PopSynth.bat', which runs each of the components of PopFor in turn. The batch file is run by giving the following argument:

PopSynth.bat %1

where: %1 is the scenario name, for example '2006_base'.

References to the scenario name (the %1 argument) have been highlighted in yellow throughout this section.

The batch file for the Population Synthesiser was written and tested during spring 2011. During late summer 2011, the functionality of the ALOGIT software was extended so that 'environment variables' could be specified allowing directory paths to be given as batch file arguments. The BTS may wish to extend the batch file supplied for the Population Synthesiser to work with environment variables once it has installed the Population Synthesiser on its machines.

For PopSynth.bat to be run, it is necessary to have input files stored in a directory named as:

Inputs_%1

The directory structure used for PopFor is detailed in the following table.

`\\ProtoSam_2006base`	Directory for PopFor system
`\\ProtoSam_2006base\Inputs_%1`	Input data that varies with scenario
`\\data`	Input data fixed for all scenarios
`\\ProtoSam_2006base\%1`	Output directory for scenario

The following table documents the batch file, with the batch file code presented on the left hand side, and an explanation of each step on the right hand side.

`rem Sydney Population Synthesiser`	
`echo off`	

```
rem create scenario and output directories
```

```
md %1
cd %1
md logfiles
md output
cd..
```

1 Create scenario folder

create directories for the various output files

```
copy Inputs_%1\Targets_%1_A1.txt Targ_1.txt
copy Inputs_%1\Targets_%1_A2.txt Targ_2.txt
copy Inputs_%1\Targets_%1_A3.txt Targ_3.txt
copy Inputs_%1\Targets_%1_A4.txt Targ_4.txt
```

2 Copy the QUAD inputs: zonal targets by area type

```
copy Inputs_%1\CatNm_%1_1.txt CatNm_1.txt
copy Inputs_%1\CatNm_%1_2.txt CatNm_2.txt
copy Inputs_%1\CatNm_%1_3.txt CatNm_3.txt
copy Inputs_%1\CatNm_%1_4.txt CatNm_4.txt
```

target averages by area type

```
copy Inputs_%1\Apriori_%1_1.txt Apriori_1.txt
copy Inputs_%1\Apriori_%1_2.txt Apriori_2.txt
copy Inputs_%1\Apriori_%1_3.txt Apriori_3.txt
copy Inputs_%1\Apriori_%1_4.txt Apriori_4.txt
```

apriori distributions by area type

```
QUAD.exe QUAD_1.ctl
QUAD.exe QUAD_2.ctl
QUAD.exe QUAD_3.ctl
QUAD.exe QUAD_4.ctl
```

3 Run QUAD, separately by area type

```
del Targ_1.txt
del Targ_2.txt
del Targ_3.txt
del Targ_4.txt
```

4 Tidy the main directory:

delete target files

```
del CatNm_1.txt
del CatNm_2.txt
del CatNm_3.txt
del CatNm_4.txt
```

delete target averages

del Apriori_1.txt del Apriori_2.txt del Apriori_3.txt del Apriori_4.txt	delete *apriori* distributions
copy Inputs_%1\HMZONE.txt alo4ec.exe merge_v6.alo del HMZONE.txt	5 Merge QUAD outputs by area type onto a single file
copy QUAD_1.log %1\logfiles\QUAD_1.log copy QUADP_1.dat %1\output\QUADP_1.dat del QUAD_1.log del QUADP_1.dat copy QUAD_2.log %1\logfiles\QUAD_2.log copy QUADP_2.dat %1\output\QUADP_2dat del QUAD_2.log del QUADP_2.dat copy QUAD_3.log %1\logfiles\QUAD_3.log copy QUADP_3.dat %1\output\QUADP_3.dat del QUAD_3.log del QUADP_3.dat copy QUAD_4.log %1\logfiles\QUAD_4.log copy QUADP_4.dat %1\output\QUADP_4.dat del QUAD_4.log del QUADP_4.dat copy merge_v6.log %1\logfiles\merge_v6.log del merge_v6.log copy QUADP_%1.dat QUADP.dat echo off	6 Copy across QUAD outputs by area type, then delete from the root folder
copy Inputs_%1\CATadj.dat	7 Copy across input files

copy Inputs_%1\lichold.dat copy Inputs_%1\income.inc		required for the licence and car ownership models
alo4ec.exe LICADJ2.alo alo4ec.exe LICPROB.alo alo4ec.exe COMPCARO_%1.alo alo4ec.exe TOTCARO_%1.alo alo4ec.exe CATSEG_%1.alo	8	Run the licence, car ownership and category segment processes
del CATadj.dat del lichold.dat del income.inc copy LICADJ2.F12 %1\logfiles\LICADJ2.F12 copy LICADJ2.log %1\logfiles\LICADJ2.log copy LICPROB.log %1\logfiles\LICPROB.log copy COMPCARO_%1.log %1\logfiles\COMPCARO_%1.log copy TOTCARO_%1.log %1\logfiles\TOTCARO_%1.log copy CATSEG_%1.log %1\logfiles\CATSEG_%1.log copy LICPROB_v9.dat %1\Output\LICPROB_v9.dat copy COMPCARO_%1_v4.dat %1\output\COMPCARO_%1_v4.dat copy TOTCARO_%1_v3.dat %1\output\TOTCARO_%1_v3.dat copy CATSEG.dat %1\output\CATSEG.dat del LICPROB_v9.dat del COMPCARO_%1_v4.dat del TOTCARO_%1_v3.dat del LICPROB.log del COMPCAR_%1.log del TOTCARO_%1.log del CATSEG_%1.log copy Inputs_%1\tcarsaext.txt copy Inputs_%1\pivotsegment.txt copy Inputs_%1\a2toa3.txt	9	Tidy and copy outputs across to scenario folder

copy Inputs_%1\Calib.dat	
Accum4.exe ACC2Work.CTL alo4ec.exe Post_Accum_WRK.alo copy Work.log %1\logfiles\Work.log copy COM_FREQ.dat %1\COM_FREQ.dat copy COMFREQ.dat %1\COMFREQ.dat copy Post_Accum_WRK.log %1\del Post_Accum_WRK.log del Work.log del COM_FREQ.dat del COMFREQ.dat del Post_Accum_WRK.log Accum4.exe ACC2Bus.CTL alo4ec.exe Post_Accum_BUS.alo copy Business.log %1\logfiles\Business.log copy BUS_FREQ.dat %1\BUS_FREQ.dat copy BUSFREQ.dat %1\BUSFREQ.dat copy Post_Accum_BUS.log %1\del Post_Accum_BUS.log del Business.log del BUS_FREQ.dat del BUSFREQ.dat del Post_Accum_BUS.log Accum4.exe ACC2SHP.CTL alo4ec.exe Post_Accum_SHP.alo copy Shopping.log %1\logfiles\Shopping.log copy SHOP_FREQ.dat %1\SHOP_FREQ.dat copy SHOPFREQ.dat %1\SHOPFREQ.dat copy Post_Accum_SHP.log %1\del Post_Accum_SHP.log del Shopping.log del SHOP_FREQ.dat del SHOPFREQ.dat del Post_Accum_SHOP.log Accum4.exe ACC2OTH.CTL alo4ec.exe Post_Accum_OTH.alo copy Other.log %1\logfiles\Other.log copy Oth_FREQ.dat %1\Oth_FREQ.dat copy OthFREQ.dat %1\OthFREQ.dat copy Post_Accum_OTH.log %1\del Post_Accum_OTH.log del Other.log	10 Run ACCUM for each home based tour purpose

```
del OTH_FREQ.dat
del OTHFREQ.dat
del Post_Accum_OTH.log

Accum4.exe ACC2PRIM.CTL
alo4ec.exe Post_Accum_PRIM.alo
copy Primary.log %1\logfiles\Primary.log
copy PRIM_FREQ.dat %1\PRIM_FREQ.dat
copy PRIMFREQ.dat %1\PRIMFREQ.dat
copy Post_Accum_PRIM.log %1\del
Post_Accum_PRIM.log
del Primary.log
del PRIM_FREQ.dat
del PRIMFREQ.dat
del Post_Accum_PRIM.log

Accum4.exe ACC2SEC.CTL
alo4ec.exe Post_Accum_SEC.alo
copy Secondary.log %1\logfiles\Secondary.log
copy SEC_FREQ.dat %1\SEC_FREQ.dat
copy SECFREQ.dat %1\SECFREQ.dat
copy Post_Accum_SEC.log %1\del Post_Accum_SEC.log
del Secondary.log
del SEC_FREQ.dat
del SECFREQ.dat
del Post_Accum_SEC.log

Accum4.exe ACC2TER.CTL
alo4ec.exe Post_Accum_TER.alo
copy Tertiary.log %1\logfiles\Tertiary.log
copy TER_FREQ.dat %1\TER_FREQ.dat
copy TERFREQ.dat %1\TERFREQ.dat
copy Post_Accum_TER.log %1\del Post_Accum_TER.log
del Tertiary.log
del TER_FREQ.dat
del TERFREQ.dat
del Post_Accum_TER.log
```

11 Final tidy of root directory

```
del QUADP.dat
del tcarsaext.txt
del pivotsegment.txt
del a2toa3.txt
```

```
del Calib.dat
```

Input files

The following sections summarise the input files for each of the of the model components. Table are presented that detail:

- the file name;
- the date the file was created; and
- a brief description of the file.

Any file that has %1 in the file name is scenario specific, and therefore needs to be updated for future runs. For these files the dates given are for the 2006 base year versions of the files. In some cases, the input file to a process is the output file from an earlier process in the Population Synthesiser; in these cases the date is given as 'when run'.

QUAD

Filename	Date	Description
QUAD_1.ctl	14/06/11	QUAD control file, area type 1
QUAD_2.ctl	14/06/11	QUAD control file, area type 2
QUAD_3.ctl	14/06/11	QUAD control file, area type 3
QUAD_4.ctl	14/06/11	QUAD control file, area type 4
Targets_%1_A1.txt	14/06/11	Zonal targets, area type 1
Targets_%1_A2.txt	14/06/11	Zonal targets, area type 2
Targets_%1_A3.txt	14/06/11	Zonal targets, area type 3
Targets_%1_A4.txt	14/06/11	Zonal targets, area type 4
CatNm_%1_1.txt	02/03/11	Target averages, area type 1
CatNm_%1_2.txt	02/03/11	Target averages, area type 2
CatNm_%1_3.txt	02/03/11	Target averages, area type 3
CatNm_%1_4.txt	02/03/11	Target averages, area type 4
Apriori_%1_1_v2.txt	02/03/11	*Apriori* distribution, area type 1
Apriori_%1_2_v2.txt	02/03/11	*Apriori* distribution, area type 2
Apriori_%1_3_v2.txt	02/03/11	*Apriori* distribution, area type 3
Apriori_%1_4_v2.txt	02/03/11	*Apriori* distribution, area type 4

Licence model

Filename	Date	Description
lichold.dat	13/12/10	Licence holding rates by age-gender cohort
HPPR_2006base_V17.F12	11/11/09	Head and partner model parameters
OTPR_2006base_V17.F12	04/01/11	Other adults model parameters
PROTO_FILE_v16.dat	13/05/11	Prototypical person sample
CATadj.dat	?	Allows increases in person weights by CAT
welf_fact.inc	?	Welfare increase to incomes

Car ownership model

Filename	Date	Description
cm_2006base_v25.F12	09/02/11	Company car ownership model parameters
hhcar_2006base_v35.F12	09/02/11	Total car ownership model parameters
licprob_v11.dat	when run	Output file from licence model
Interim_parking_cost.txt	21/09/09	Zonal parking costs in centre areas
welf_fact.inc	?	Welfare increase to incomes

Category segments

Filename	Date	Description
totcaro_2006_v3.dat	when run	Output from licence and car ownership models

ACCUM

Filename	Date	Description
catseg.dat	when run	Output from category segment process
ACC2Work.CTL	09/02/11	ACCUM control file, commute
ACC2Bus.CTL	09/02/11	ACCUM control file, home–business
ACC2SHP.CTL	09/02/11	ACCUM control file, home–shop
ACC2OTH.CTL	09/02/11	ACCUM control file, home–other travel
ACC2PRIM.CTL	09/02/11	ACCUM control file, home–primary educ.
ACC2SEC.CTL	09/02/11	ACCUM control file, home–secondary educ.
ACC2TER.CTL	09/02/11	ACCUM control file, home–tertiary educ.

Output files

The output files produced by the Population Synthesiser are summarised in the following tables.

QUAD

Filename	Description
QUAD_1.log	QUAD log file, area type 1
QUAD_2.log	QUAD log file, area type 2
QUAD_3.ctl	QUAD log file, area type 3
QUAD_4.ctl	QUAD log file, area type 4
QUAD_1.dat	QUAD expansion factors by zone and CAT, area type 1
QUAD_2.dat	QUAD expansion factors by zone and CAT, area type 2
QUAD_3.dat	QUAD expansion factors by zone and CAT, area type 3
QUAD_4.dat	QUAD expansion factors by zone and CAT, area type 4
merge_v6.log	Log file for merge process

Licence model

Filename	Description
LIDADJ2.F12	Licence adjustment procedure parameters
LIDADJ2.LOG	Licence adjustment procedure log file
LICPROB.LOG	Licence holding models log file
licprob.dat	Licence holding model probabilities

Car ownership model

Filename	Description
COMPCARO.LOG	Company car ownership model log file
TOTCARO.LOG	Total car ownership model log file
COMPCARO.DAT	Company car (and licence holding) model probabilities
TOTCARO.DAT	Total car (and company car and licence holding) prob.s

Category segments

Filename	Description
catseg.dat	Category segments file classifying each person record with associated licence, company and total car ownership states into mode-destination and frequency segments