Choosing optimal stratifiers for the National Travel Survey

Shaun Scholes

Choosing optimal stratifiers for the National Travel Survey

Shaun Scholes

Prepared for the Department for Transport

July 2006

Contents

ACKNOWLEDGEMENTS

I am grateful to Kevin Pickering, Susan Purdon and Giulio Flore at the Survey Methods Unit (SMU) of the National Centre for Social Research for reading and commenting on the drafts of this report. I would like to thank Sarah Tipping (SMU) for deriving the Census based variables and advising on NTS sampling. The report has drawn extensively on the previous NTS stratification report written by Jeremy Barton of the Office for National Statistics. Finally, special thanks are due to Olivia Christophersen, Darren Williams and Barbara Noble at the Department for Transport for their technical assistance and suggestions.

EXECUTIVE SUMMARY

Using methods developed at the Office for National Statistics (ONS), the 2003 National Travel Survey (NTS) has been used in tandem with the 2001 Census to examine whether the current set of NTS postcode sector stratifiers is the most optimal available.

18 variables from the 2003 NTS were aggregated to the postcode sector level. Using postcode sector identifiers, the aggregated NTS estimates were matched to geographic classifications and 2001 Census-based information. These candidates for stratification were fitted as independent variables in a linear regression model on each of the NTS variables in turn. Each model was compared on the basis of their *adjusted* multiple coefficient of determination (R^2). This measures the percentage of variance in the dependent variable accounted for by the terms in the model. All other things being equal, the geographic classifications and Census variables explaining most of the variability in survey measures are the best choice of postcode sector stratifiers.

The first stratification variable for social surveys in Great Britain is often a regional classification. The current NTS regional stratifier (based largely on NUTS2 areas) was compared against Government Office Region. The existing regional stratifier produced the highest adjusted R^2 for the four household/individual level variables (the smallest difference in R^2 was 3.8 percentage points). The analysis of Census-based stratifiers, therefore, was conducted having taken into account the 40 NTS regions.

Optimal stratifiers for one variable often turn out to be sub-optimal for others because the correlation between survey variables and stratification factors vary for each survey measure. As a result, the variables chosen are usually those which appear in several optimal models and where there is a strong intuitive reason for their strength as stratifiers.

Car ownership measures from the 2001 Census featured in eight of the 18 models, and were the best performing variable in six cases. For example, areas with a higher percentage of households having access to two or more cars/vans (CARS2PLP) were associated with a lower percentage of households having access to a bus service that ran at least every 15 minutes during the day, and an increase in stage and trip distance. The evidence, therefore, suggests retaining CARS0P (the percentage of households with no car/van) as the second stratifier.

Census measures related to household type (e.g. one person households) and distance to work (e.g. the percentage of persons with a distance to work between 2-5 kilometres) were more strongly correlated to the NTS measures than population density (the third stratifier currently used). Population density appeared in only two of the 18 models. All other things being equal, therefore, a more optimal third stratifier may be available.

The Department for Transport, however, have stressed the usefulness of population density as a stratifier to ensure a balanced sample of households living in urban and rural areas. Population density, therefore, may be preferred as the third stratifier as it serves an important *general purpose*. In addition, no variable was found to perform consistently better than others in predicting the NTS variables.

1 INTRODUCTION

The National Travel Survey (NTS) provides up-to-date and regular information about personal travel within Great Britain and monitors trends in travel behaviour. Since January 2002, the Department for Transport (DfT) has commissioned the National Centre for Social Research (NatCen) as the contractor for the NTS. NatCen is responsible for questionnaire development, sample selection, data collection and editing, and data file production. The DfT is responsible for building the database and data analysis, publication and archiving.

The current set of regional and Census-based stratifiers on the NTS were introduced in 2002. They consist of:

- NTS regional breakdown based on NUTS2 areas;
- car ownership (percentage of households with no car); and
- population density (persons per hectare).

Each stratifier is defined at the postcode sector level. These stratifiers were chosen as a result of empirical work by the Methods and Sampling Unit of the Office for National Statistics (Barton,1996). This report updates that work by using the 2003 NTS data in tandem with information from the 2001 Census to consider whether the current set of NTS stratifiers is optimal, and if not, whether a new set would achieve enough gains in terms of increasing the precision of estimates to warrant replacing the existing set. The 2003 survey has been chosen for its larger sample size and its proximity to the 2001 Census.

1.1 Stratification in sample surveys

In survey sampling explicit stratification involves the division of *all* Primary Sampling Units (PSUs) such as postcode sectors into sub-groups or strata from which independent samples are taken. Implicit stratification, in contrast, involves systematically selecting PSUs from an ordered list rather than grouping them. NatCen typically uses implicit stratification or a mixture of explicit and implicit methods.

Stratifying the PSUs ensures that a representative sample will be drawn with respect to the set of stratifiers used. It increases the precision of survey estimates (relative to a sample selected without stratification) *if* there is a correlation between stratification factors and survey variables (Korovessis,2001). Because stratification increases the effective sample size, choosing the most powerful set of stratifiers can lead to increased sample efficiency.

1.2 Sampling and stratification of the 2003 National Travel Survey

For the NTS a two-stage sample of addresses is used where addresses are sampled from a stratified sample of postcode sectors. The 2003 NTS was based on a random sample of 15,048 private households, drawn from the Postcode Address File (PAF). The sample was designed to ensure that the addresses for each quarter were representative of the total Great Britain population.

From 2002 onwards the NTS has used a 'quasi-panel' design. According to this design, half the PSUs in a given year's sample are retained (using random sampling) for the next year's sample and the other half are replaced. This has the effect of reducing the variance of estimates of year-on-year change. (342 of the PSUs selected for the 2002 NTS were retained for the 2003 sample, supplemented with 342 new PSUs). The PSUs carried over from the 2002 NTS were excluded from the 2003 sample frame, so they could not appear twice in the sample.

For the 2003 sample a list of all postcode sectors in Britain was generated, excluding those in the Scottish Islands and the Isles of Scilly. Sectors south of the Caledonian Canal with less than 500 delivery points and sectors north of the Caledonian Canal with less than 250 delivery points were grouped with an adjacent sector. Grouped sectors were then treated as one PSU.

The list of postcode sectors in Great Britain was stratified using a regional variable and Census measures of car ownership and population density. This is done in order to increase the precision of the sample and to ensure that the different strata in the population are correctly represented. Random samples of PSUs were then selected within each stratum.

The regional strata for Great Britain is based on the NUTS2 areas, grouped in a few cases where single areas are too small. NUTS or Nomenclature of Units for Territorial Statistics is a European-wide geographical classification developed by the European Office for Statistics (Eurostat). NUTS2 roughly relates to counties or groups of counties in England, and groups of unitary authorities or council areas in Scotland and Wales. The 40 regions currently used as the first NTS stratifier are shown in Table 1.

Within each NTS region, postcode sectors were listed in increasing order of the percentage of households with no car (according to the *1991 Census*). Cut-off points were then drawn approximately one third and two thirds (in terms of delivery points) down the ordered list, to create three roughly equal-sized bands. Within each of the 120 bands thus created (40 regions by 3 car ownership bands), sectors were listed in order of population density (persons per hectare). 342 postcode sectors were then systematically selected with probability proportional to delivery point count (after expansion by the Multiple Occupancy Count in Scotland). Differential sampling fractions were used in Inner London, Outer London and the rest of Great Britain in order to oversample London (with the aim of achieving responding sample sizes that reflect the regional distribution without the need for corrective weighting). These 342 'new' sectors were then added to the 342 `quasi-panel'

sectors carried over from NTS 2002 to make the final sample of 684 postcode sectors for NTS 2003.

Table 1 NTS regional stratification variable

	England	Government Office Region
1	Inner London - East	London
2	Inner London - West	London
3	Outer London - East and North East	London
4	Outer London - South	London
5	Outer London - West and North West	London
6	Devon and Cornwall	South West
7	North Somerset, North East Somerset, Bath, Somerset and Dorset	South West
8	Bristol, South Gloucestershire, Gloucestershire and Wiltshire	South West
9	Oxfordshire, Buckinghamshire and Berkshire	South East
10	Hampshire and Isle of Wight	South East
11	Kent	South East
12	West Sussex and East Sussex	South East
13	Surrey	South East
14	Essex	East of England
15	Cambridgeshire, Suffolk and Norfolk	East of England
16	Hertfordshire and Bedfordshire	East of England
17	Leicestershire, Lincolnshire and Northamptonshire	East Midlands
18	Warwickshire and Hereford & Worcester	West Midlands
19	West Midlands	West Midlands
20	Shropshire and Staffordshire	West Midlands
21	Nottinghamshire and Derbyshire	East Midlands
22	Cheshire	North West
23	Merseyside	North West
24	Greater Manchester	North West
25	Lancashire and Cumbria	North West
26	South Yorkshire	Yorkshire and The Humber
27	West Yorkshire	Yorkshire and The Humber
28	North Yorkshire and Humberside	Yorkshire and The Humber
29	Cleveland, County Durham and Northumberland	North East
30	Tyne & Wear	North East

	Wales	Government Office Region
31	Anglesey, Gwynedd, Conwy, Denbighshire, Flintshire, Wrexham, Powys, Ceredigion	Wales
32	Carmarthenshire, Neath Port Talbot, Pembrokeshire, Swansea	Wales
33	Blaenau Gwent, Caerphilly, Monmouthshire, Newport, Torfaen	Wales
34	Bridgend, Cardiff, Merthyr Tydfil, Rhondda Cynon Taff, Vale of Glamorgan	Wales
	Scotland	**Government Office Region**
35	Grampian, Highland, Argyll & Bute	Scotland
36	Tayside, Fife and Central	Scotland
37	Edinburgh, Lothians and Borders	Scotland
38	Glasgow and Dunbartonshire	Scotland
39	Lanarkshire, Renfrewshire and Inverclyde	Scotland
40	Ayrshire and Dumfries & Galloway	Scotland

2 METHODOLOGY

The Office for National Statistics (ONS) have conducted a number of empirical trials aimed at optimising the choice of stratifiers for the Family Resources Survey, General Household Survey and National Travel Survey. The methodology used can be summarised in the following main steps.

2.1 Choosing survey variables and aggregating to PSU level

A range of dependent variables from the survey in question are aggregated to the postcode sector level. For example, a variable indicating whether an adult was in possession of a full car driving licence can be aggregated to estimate for each sampled PSU the percentage of adults holding a full driving licence.

2.2 Regional and Census-based independent variables

An analysis dataset at the PSU level is compiled by linking (via postcode sector identifiers) the aggregated survey estimates described above to geographic classifications such as Government Office Region and Census-based information. These variables encompass both the current and possible alternative PSU level stratification factors.

2.3 Initial regression models

The basic approach to choosing new stratifiers involves using the analysis dataset to fit the possible stratification factors as independent variables in a linear regression model, on each of the key survey dependent variables in turn (Bruce,1993;Barton,1996;Insalaco,2000). The stratifying variables are initially analysed as continuous variables (except for region), but are

then grouped so that different combinations of bands (e.g. tertiles or quartiles) can be tested. The banding serves to replicate how the second stratifier is actually used when selecting PSUs from the sampling frame.

Each model is then compared on the basis of their *adjusted* multiple coefficient of determination (R^2). This measures the percentage of variance in the dependent variable accounted for by the terms in the regression model. The adjusted R^2 is a goodness-of-fit measure in multiple regression analysis that penalises the inclusion of additional independent variables by using a degrees of freedom adjustment in estimating the error variance (Wooldridge,2003). This makes the adjusted R^2 a useful measure for comparing models based on a different number of independent variables.

For each survey variable, the regressions are performed in a stepwise procedure. That is, once the first stratifier is decided upon, the second is chosen with the first factor already included in the model. Similarly, the third stratification factor is chosen with the first and second factors already included in the model.

All things being equal, the two Census variables appearing most often in the final optimal models (i.e. those which explain most of the variability in survey measures), in combination with the chosen regional stratifier, are, in theory, the best choice and so are chosen as the second and third stratifiers.

2.4 Evaluate increases in precision

To examine the extent of any gain in precision, the adjusted R^2's from two models (one containing the proposed stratifiers and one existing stratifiers) are compared. The percentage gain in precision (if any) achieved by the proposed stratifiers can then be computed by the following formula:

$$\%\ gain\ in\ precision = \left[\frac{R^2_{new stratifiers} - R^2_{old\ stratifiers}}{1 - R^2_{old\ stratifiers}} \right] \times 100$$

The method set out in this section has proved to be a fast and efficient way of looking at the effect of many potential Census-based PSU level stratifiers on a number of key variables across a range of surveys (Barton,1996;Insalaco,2000). It is important, however, to bear two limitations of the method in mind.

First, both the regression modelling to choose new stratifiers and evaluation of the potential gains in precision tend in practice to be conducted on the *same* survey dataset. This may lead to the problem of 'over-fitting' the data (i.e. building models that were optimal for the dataset at hand but perform very poorly on future survey data). The gains in precision, therefore, may be too optimistic.

Second, the methodology has been criticised for giving a misleading impression in showing how the stratification of PSUs impacts on the variance of household-based estimates. Any potential improvements to statistical efficiency brought about by stratification affect only the

between-PSU component of sampling variance (Thomas and Pickering,2003). The between-PSU component of sampling variance is minor compared to the variance *within PSUs*. Thus, for example, if a reduction in variance of 15% is predicted for a new set of PSU stratifiers as compared with the existing set, but the between-PSU component accounts for only 10% of *total* between-households variance, then the actual reduction in the variance of household-based estimates is only 1.5% (Thomas and Pickering,2003,p.22).

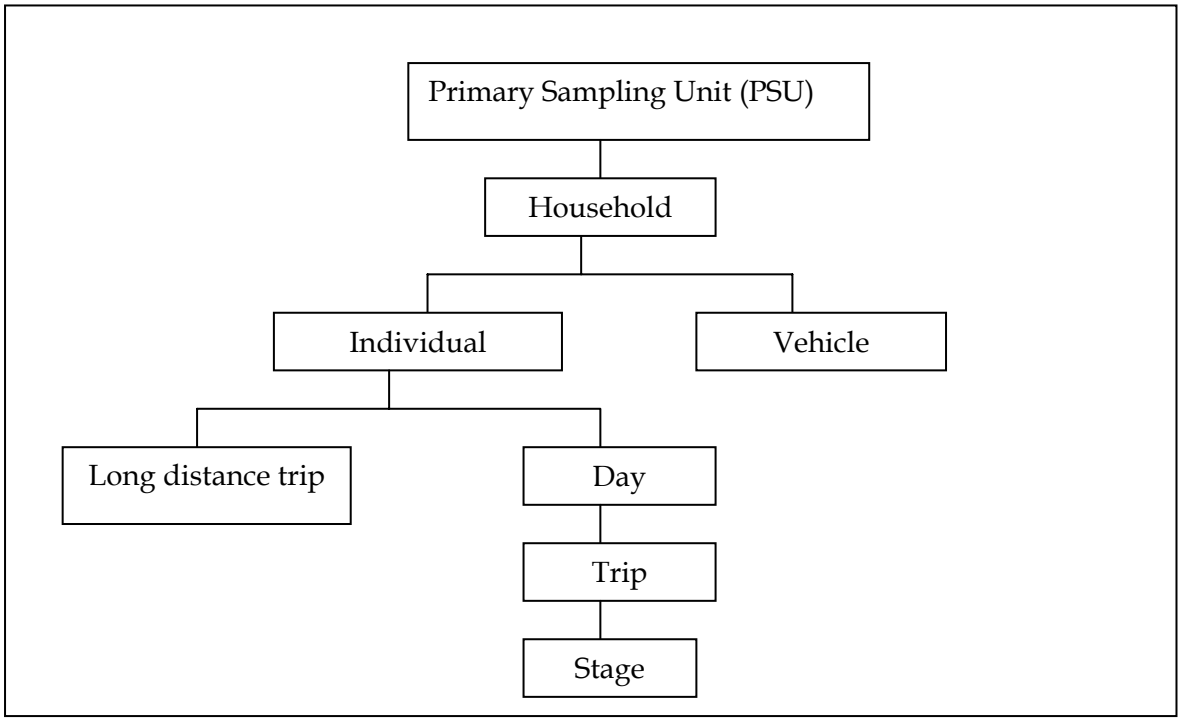
Finally, it is worth mentioning that, in practice, the choice of stratifiers can often owe something to considerations other than maximising the precision of estimates. For example, survey users may wish to be assured that the proportion of the sample assigned to certain groups is controlled, irrespective of whether other variables may perform marginally better in predicting the survey variables.

3 ANALYSIS AND RESULTS

3.1 NTS key variables

The NTS gathers information about several different aspects of travel including: purpose of travel, method of travel (walk, car, bus etc.), origin and destination of trips, time travelling and distance, as well as detailed information about households, individuals, and vehicles. The NTS data is held in an eight-level hierarchical database, as shown in Figure 1.

Figure 1 The NTS 8-level hierarchical database



As Barton (1996) notes, although the NTS has a specific subject area to investigate, it remains possible to choose a broad selection of variables related to transport (e.g. variables describing different forms of transport). Other criteria for the choice of variables in the earlier NTS stratification report were the extent of use in NTS reports and the size of the design factor. The design factor represents the ratio of the standard error of estimates from the actual sample design used to the standard error of estimates from a simple random sample of the same size. A design factor of 1.5, for example, indicates that the estimate has a standard error that is 50% as large as would have been obtained from simple random sampling (i.e. a sample where every unit in the population has an equal and independent probability of selection).

This updated stratification report largely uses the same set of NTS variables analysed by Barton (1996) together with those recommended by the DfT.

Table 2 lists the 18 NTS variables analysed in this report together with its design factor estimated from the 2003 survey. Note that the analysis was performed on the 8,258 fully responding households in 2003.[1] These households contained 19,467 individuals of all ages with information on 314,845 trips. (A trip is a one-way course of travel having a single main purpose). The trips themselves were broken down into 327,230 stages. (A trip consists of one or more stages. A new stage is defined when there is a change in the form of transport or when there is a change of vehicle requiring a separate ticket).

As Insalaco (2000,p.7) notes, the optimal stratifiers for one variable are likely to be different than for another because the correlation between survey variables and stratification factors will be different for each survey variable. Choosing optimal stratifiers, therefore, requires making a compromise between the optimal solutions for a range of variables. It is for this reason that variables were selected across different levels of the NTS database.

[1] For a household to be classed as fully co-operating, the placement interview had to be fully completed and filled in travel diaries had to be collected for all household members (for further details see Hayllar *et.al* (2005)).

Table 2 Key variables chosen from the 2003 NTS

NTS variable and level	Description	N	Estimate	Design factor[1]
Household			%	
BUS15MIN	Households with bus service every 15 minutes	8,258	33.9	1.70
BIKEINHH	Households with 1+ bikes	8,258	47.6	1.09
WALKRAIL	Households <13 minutes walk to railway station	8,258	15.6	1.93
Individual			%	
DRIVLICE	Adults with full car driving licence	15,288	69.4	1.28
Trip			%	
SHOPPING	Purpose of trip 'to' was shopping	314,845	10.7	1.86
			Mean	
J29	Overall travelling time	314,845	22.9	3.24
JOTXSC[2]	Overall trip time	314,845	25.7	3.20
JD[2]	Trip distance (in tenths of miles)	314,845	81.0	2.71
JJXSC[2]	Number of trips	314,845	1.2	2.54
JDSCHOOL[2]	School trip distance	6,087	25.7	2.98
Stage			%	
STGEWALK	Stages where mode of transport was walking	327,230	12.7	4.27
STGECAR	Stages where mode of transport was car	327,230	69.9	5.31
			Mean	
STTXSC[2]	Stage travelling time	327,230	23.9	3.12
SD[2]	Stage distance	327,230	78.0	2.71
S25	Length of stage (tenths of mile)	327,230	77.2	2.71
S36	Stage travel time	327,230	22.0	3.20
TRAVMINS	Trip travelling time (valid cases)	325,749	21.9	3.23
Vehicle			*Mean*	
V46	Annual mileage	9,264	9133.9	1.05

Notes:

[1] Both the estimates and design factors were estimated at the appropriate level of the NTS database (i.e. household, individual, trip, stage and vehicle). The design factors were estimated in STATA using a strata variable which merged pairs of adjacent PSUs within each NTS region by car ownership band. (If there was an odd number of PSUs within a stratification band a pair of PSUs was joined by an adjacent PSU). The PSU was specified as the clustering variable.

[2] These variables are weighted to account for short walks, which are only collected on the seventh day of the travel week, and exclude 'series of calls'.[2]

[2] When a respondent makes a series of short stops on a journey (e.g. when a respondent makes a long shopping trip and calls at numerous shops), the respondent or interviewer may find it difficult to subdivide the journey into shorter trips. In this case, the trips are combined into a single trip, known as a 'series of calls' trip (Hayllar *et.al*,2005,p.116).

3.1.1 Weighting

Weighting the data to correct for non-response bias will be introduced from 2006 onwards when NTS data for 2005 and previous years will be published on a weighted basis. The analyses for this stratification report have been performed on un-weighted data as at the time of writing the weights were in the process of being rigorously checked. The conclusions based on weighted and un-weighted data are unlikely to be different.

3.2 Choosing a regional classification

As mentioned in Section 1.3 the first NTS stratification factor is a regional variable based on NUTS2 areas. Table 3 describes the current (NTSREG) and alternative regional stratifier (Government Office Region) available on the sampling frame of postcode sectors regularly used by NatCen. (Standard Statistical Regions are no longer used by the DfT and so were not considered).

Table 3 Current and alternative regional stratifier

Regional variable	Number of regions	Description
NTSREG	40	Based on NUTS2 areas
GOR	12	Government Office Region for England (London split into inner and outer regions) with Wales and Scotland as single areas

Both regional variables were taken from the *current* version of the Postcode Address File (Version 41). At the PSU level, linear regression models were fitted using NTS variables as the outcome measure and the regional variables as categorical independent factors. The adjusted R^2 for each model is given in Table 4 with the largest value for each variable shown in a shaded cell. [3]

[3] As was the case in the earlier NTS stratification report, the R^2 actually represents a biased estimate of the "true" R^2 as the effect of sampling households within PSUs inflates the variance between area means (i.e. a portion of the area level variance actually represents variance between households within the same area). Although this bias affects the final estimates of precision the comparison between different models is not affected (Barton,1996).

Table 4 Adjusted R^2 for each regional classification by NTS variable

	Regional classification	
NTS variable	*NTSREG*	*GOR*
BUS15MIN	38.9%	32.4%
BIKEINHH	16.1%	10.4%
WALKRAIL	18.0%	14.2%
DRIVLICE	15.6%	10.8%
SHOPPING	1.8%	2.7%
J29	28.6%	28.9%
JOTXSC	34.1%	34.2%
JD	6.1%	4.9%
JJXSC	3.6%	3.9%
JDSCHOOL	0.3%	0.0%
STGEWALK	4.4%	3.9%
STGECAR	38.6%	36.5%
STTXSC	20.2%	20.3%
SD	8.4%	6.7%
S25	8.6%	6.9%
S36	16.8%	17.2%
TRAVMINS	16.9%	17.3%
V46	2.5%	3.2%

The results show that the current NTS regional stratifier produces the highest adjusted R^2 for 10 of the 18 survey variables, including the four household/individual level variables. The higher the value of R^2, the better the independent variables would perform as stratifiers (Bruce,1993). For example, NTSREG explained 8.6% of the variation in length of stage (S25) compared to 6.9% by GOR (after imposing a penalty for adding additional terms to the model). Where GOR did perform better than NTSREG the difference was less than one percentage point. These results suggest that the current regional stratifier is the most optimal one available. Hence NTSREG was employed in all subsequent analyses as the first stratifier.

3.3 Choosing Census-based stratifiers

The second and third stratifiers for the 2003 NTS (proportion of households with no car and population density) were taken from the 1991 Census as the 2001 results were not available at the time of sampling. The 2001 Census provides a range of variables that could be used to stratify postcode sectors. 40 variables have been chosen as potential candidates for stratification. These are shown in Table 5.[4] The variables have been selected on the grounds that they measure aspects of travel (e.g. mode of travel to work, distance to work and car ownership) and that they were used in the earlier NTS stratification report. In addition a number of area-level variables employed in analyses of non-response to the NTS were also considered (see Pickering *et.al*, 2005).

[4] Five postcode sectors selected for the 2003 NTS were not represented in the current version of the Postcode Address File. Their values on the 2001 Census variables were imputed using the average values for postcode sectors in the same district.

Table 5 Potential stratifying variables available from the 2001 Census

Variable	Description
POPDENS	Population density (persons per hectare)
OWNP	% Owner occupier households
LARENT	% Households rented from council
PRIVREN	% Private renting households
NSSEC12	% Household Reference Persons (HRP) in NS-SEC categories 1 and 2[1]
NONWHITE	% Persons non-white
PENSION	% Persons who are pensioners (men >65, women >60)
OVER75	% Persons aged 75 or over
CARS0P	% Households with no car/van
CARS2PLP	% Households with 2 or more cars/vans
CARSAVP	Average number of cars per household
SEMIDETP	% Households semi-detached
LLTIYP	% Persons with limiting long-term illness
EMPDP	% Persons aged 16-74 economically active
UNEMP	% Persons aged 16-74 unemployed
ECINP	% Persons aged 16-74 economically inactive
HOMEKM	% Persons working mainly at home[2]
LESS2KM	% Persons with distance to work less than 2 km[2]
TWO5KM	% Persons with distance to work between 2-5 km[2]
FIVE10KM	% Persons with distance to work between 5-10 km[2]
TEN20KM	% Persons with distance to work between 10-20 km[2]
TWENTYKM	% Persons with distance to work greater than 20 km[2]
WRKTRAIN	% Persons travel to work by train[2,3]
WRKCAR	% Persons travel to work by car/van[2,4]
MARITAL1	% Adults lone parents
MARITAL2	% Adults married/cohabiting
MARITAL3	% Adults other marital status – including single
NOKIDSP	% Households with no children
HHSIZE1P	% Households with 1 person
HHSIZE2P	% Households with 2 persons
HHSIZE3P	% Households with 3 persons
HHSIZE4P	% Households with 4 or more persons
HHSIZAV	Average household size
HP1634P	% Households with HRP aged 16-34
HP3544P	% Households with HRP aged 35-44
HP4554P	% Households with HRP aged 45-54
HP5564P	% Households with HRP aged 55-64
HP65PLP	% Households with HRP aged 65 or over
MUNEMP	% Males who are unemployed
FUNEMP	% Females who are unemployed

Notes:

[1] Higher and lower managerial and professional occupations.

[2] All persons resident in area aged 16-74 (England and Wales). All persons resident in area (Scotland). Scottish counts are for persons working or studying.

[3] Includes underground, metro, light rail and tram.

[4] As driver or passenger.

Stepwise forward selection regressions were carried out to examine which Census variables (fitted as continuous terms) were most highly correlated with each NTS variable, having taken account of the 40 NTS regions.[5] The variables that were selected at each step (up to a maximum of four) are shown in Table 6. (If additional variables did not add significantly to the explanatory power of the model, fewer than four were recorded).

(It is worth noting that while the variables were those selected on the basis of contributing most to an increase in R^2, the actual difference in the magnitude of the increase between the highest and the next several was often relatively small. That is, similar R^2 values could often be attained by various combinations of variables).

Table 6 Variables entered on forward stepwise procedures

	2001 Census variables after NTS region already included			
NTS variable	*Step 1*	*Step 2*	*Step 3*	*Step 4*
BUS15MIN	CARS2PLP	TWO5KM	LLTIYP	POPDENS
BIKEINHH	CARSAVP	NSSEC12	-	-
WALKRAIL	WRKTRAIN	HHSIZ1P	TWO5KM	HOMEKM
DRIVLICE	CARS0P	HHSIZ1P	NOKIDSP	-
SHOPPING	WRKTRAIN	HHSIZ2P	CARSAVP	POPDENS
J29	PRIVREN	HOMEKM	-	-
JOTXSC	PRIVREN	TWO5KM	OWNP	NSSEC12
JD	CARS2PLP	TWO5KM	HP1634P	HHSIZAV
JJXSC	WRKCAR	ECINP	HHSIZ2P	MARIT1
JDSCHOOL	HOMEKM	TWENTYKM	-	-
STGEWALK	WRKCAR	LESS2KM	MARITAL3	HP65PLP
STGECAR	WRKCAR	HP1634P	CARS0P	ECINP
STTXSC	PRIVREN	TWO5KM	OWNP	NSSEC12
SD	CARS2PLP	TWO5KM	HHSIZAV	HP1634P
S25	CARS2PLP	TWO5KM	HHSIZAV	OVER75
S36	HOMEKM	HP1634P	LESS2KM	TWO5KM
TRAVMINS	HOMEKM	HP1634P	LESS2KM	-
V46	TWENTYKM	MUNEMP	PENSION	HHSIZ3P

First, the results reaffirm the strong correlation between NTS variables and Census measures of car ownership, including variables not directly related to car travel (e.g. the proportion of households with one or more bikes). Car ownership measures featured in eight of the 18 models, and were the best performing variable in six cases. These were spread across different levels of the NTS hierarchical database (household, individual, trips and stage).

[5] Stepwise regression is a method for selecting the independent variables for a regression model (Insalaco,2000). In this case, an initial regression of the NTS survey variable and NTS region (as a categorical variable) was set and a list of variables was specified. The variable from the list that was the most statistically significant for the initial regression was added to the model as an additional independent variable. The procedure then examined which of the remaining variables from the list was the most statistically significant for the regression and included it as an additional variable. The procedure continued until there were no variables left in the list that were statistically significant for the model.

For example, areas with a higher percentage of households having access to two or more cars/vans (CARS2PLP) were associated with a lower percentage of households having access to a bus service than ran at least every 15 minutes during the day (BUS15MIN). CARS2PLP was also associated with an increase in stage distance (SD), length of stage in tenths of mile (S25) and trip distance (JD). This analysis, therefore, confirms that a car ownership measure should continue as the second stratifier. It seems sensible that CARS0P, the percentage of households with no car/van, should be retained as the measure of car ownership to ensure no loss in continuity.[6]

Second, population density (currently the third NTS stratifier) appears in only two of the 18 models. The analysis suggests that Census measures related to either household type (e.g. household size or age of the Household Reference Person) or travel to work (e.g. the percentage of persons working mainly at home or with a distance to work between 2-5 kilometres) may be a more optimal choice. For example, areas with a higher percentage of persons working mainly at home (HOMEKM) were associated with an increase in school trip distance (JDSCHOOL) and stage travel time (S36).

The DfT, however, have emphasised the importance of retaining population density as a stratifier to ensure a balanced sample of households living in urban and rural areas. Although POPDENS is a *proxy* measure of urbanity it is a continuous measure at the postcode sector level of geography. (This is in contrast to the eight or four-fold rural and urban classification available for other levels of geography including Census Output Areas, Census Area Statistics wards and Local Authorities). There is a case, therefore, for retaining POPDENS as the third stratifier *despite* evidence suggesting that other Census variables perform better in predicting the NTS variables.

Before proceeding to assess further the possible candidates for the third stratifier the next step involved examining any potential gains in precision from increasing the number of car ownership bands within each NTS region from three to four.

3.4 Selecting appropriate number of bands for car ownership

The NTS currently allows for 40 regions and three car ownership bands within each of these (120 bands in total). Given the increase in the NTS sample size from 2002 onwards and the retainment of half the PSUs in a given year's sample for the following year's sample the possibility of increasing the number of car ownership bands within each NTS region from three to four was investigated. The two models tested using the NTS 2003 data and the 2001 Census measure of car ownership were:

Model 1: NTSREG × CARS0P (3 bands)
Model 2: NTSREG × CARS0P (4 bands)

The adjusted R^2 for each model is given in Table 7.

[6] CARS0P and CARS2PLP are strongly associated (a correlation coefficient of -0.9 using the complete list of PSUs) and so the strong correlation between CARS2PLP and NTS variables such as BUS15MIN, JD, SD and S25 can be expected to be mirrored in practice by using CARS0P as the car ownership measure.

Table 7 Adjusted R^2 for models 1 and 2 by NTS variable

NTS variable	Model 1	Model 2
BUS15MIN	50.6%	53.1%
BIKEINHH	19.4%	18.6%
WALKRAIL	23.0%	22.6%
DRIVLICE	42.6%	42.1%
SHOPPING	0.8%	0.7%
J29	28.9%	33.4%
JOTXSC	34.7%	39.6%
JD	13.6%	16.1%
JJXSC	9.8%	9.7%
JDSCHOOL	6.9%	7.0%
STGEWALK	14.9%	15.0%
STGECAR	53.9%	55.5%
STTXSC	20.7%	25.4%
SD	16.7%	19.3%
S25	17.2%	19.7%
S36	18.3%	22.7%
TRAVMINS	18.4%	23.2%
V46	1.0%	1.4%

Moving from three to four car ownership bands within each NTS region increased the amount of variance explained in nine of the 18 survey variables. In these cases the increase in adjusted R^2 was greater than two percentage points. In four cases, the adjusted R^2 decreased but the size of the actual difference was small (less than one percentage point).[7]

One drawback of increasing the number of car ownership bands would be to weaken the effect of the third stratifier. All other things being equal, increasing the number of NTSREG × CARS0P bands from 120 to 160 would result in a less finely tuned selection of PSUs (in terms of population density) within each band. For this reason it is recommended to continue to use three car ownership bands within each NTS region.

3.5 Choosing the final stratifier

To further examine the case for replacing POPDENS as the third stratifier the stepwise analyses were repeated taking account of both NTS region and car ownership. The Census measures were fitted as continuous variables (i.e. in ungrouped form). The results for a maximum of three steps are shown in Table 8.[8]

[7] A similar mixed pattern was found in the earlier NTS stratification report (see columns 5 and 8 of Table 5 in Barton,1996).

[8] CARS2PLP and CARSAVP were not candidates for inclusion in the forward stepwise models as CARS0P was already included.

Table 8 Variables entered on forward stepwise procedures

	2001 Census variables after NTS region and CARS0P (3 bands) already included		
NTS variable	*Step 1*	*Step 2*	*Step 3*
BUS15MIN	POPDENS	TWO5KM	WRKCAR
BIKEINHH	MUNEMP	HOMEKM	-
WALKRAIL	WRKTRAIN	TWO5KM	HHSIZ1P
DRIVLICE	NOKIDSP	WRKCAR	NSSEC12
SHOPPING	POPDENS	HHSIZ2P	WRKTRAIN
J29	NSSEC12	TWO5KM	OWNP
JOTXSC	PRIVREN	TWO5KM	NSSEC12
JD	TWENTYKM	UNEMP	TWO5KM
JJXSC	WRKCAR	PENSION	HHSIZ1P
JDSCHOOL	HOMEKM	POPDENS	-
STGEWALK	MARIT3	LESS2KM	WRKCAR
STGECAR	WRKCAR	HP1634P	ECINP
STTXSC	PRIVREN	TWO5KM	NSSEC12
SD	TWENTYKM	UNEMP	TEN20KM
S25	UNEMP	TWENTYKM	TEN20KM
S36	NOKIDSP	LLTIYP	TWO5KM
TRAVMINS	NOKIDSP	LLTIYP	TWO5KM
V46	PRIVREN	TWENTYKM	-

The six best performing variables in rank order were: TWO5KM, TWENTYKM, WRKCAR, NOKIDSP, PRIVREN and POPDENS. These variables are shown by the shaded cells in Table 8. To further examine this subset of six variables, the adjusted R^2 for each NTS measure were compared. The results are shown in Table 9.

Table 9 Adjusted R^2 for models including TWO5KM, TWENTYKM, WRKCAR, NOKIDSP, PRIVREN and POPDENS as third NTS stratifier

	Adjusted R^2					
NTS variable	*NTSREG + CARS0P + TWO5KM*	*NTSREG + CARS0P + TWENTYKM*	*NTSREG + CARS0P + WRKCAR*	*NTSREG + CARS0P + NOKIDSP*	*NTSREG + CARS0P + PRIVREN*	*NTSREG + CARS0P + POPDENS*
BUS15MIN	56.9%	54.1%	51.6%	51.3%	50.9%	57.0%
BIKEINHH	19.2%	19.4%	19.4%	19.5%	19.2%	19.3%
WALKRAIL	25.1%	23.4%	24.7%	24.9%	26.2%	23.1%
DRIVLICE	42.8%	43.2%	43.3%	47.1%	42.7%	42.8%
SHOPPING	0.7%	0.6%	0.7%	1.4%	2.1%	2.7%
J29	29.8%	29.6%	29.5%	30.3%	30.5%	28.8%
JOTXSC	35.9%	35.6%	36.7%	35.8%	38.1%	34.7%
JD	17.4%	19.1%	13.5%	17.2%	15.1%	14.6%
JJXSC	11.0%	10.1%	17.5%	9.6%	13.5%	11.3%
JDSCHOOL	7.3%	8.6%	6.6%	7.8%	6.9%	9.6%
STGEWALK	17.1%	16.3%	24.6%	16.5%	25.0%	15.7%
STGECAR	53.9%	53.8%	60.6%	53.9%	59.6%	55.5%
STTXSC	22.0%	21.7%	22.6%	22.1%	23.3%	20.7%
SD	20.3%	21.8%	16.7%	19.9%	17.7%	17.9%
S25	20.6%	22.2%	17.2%	20.3%	18.0%	18.4%
S36	19.1%	18.9%	18.5%	19.4%	19.2%	18.2%
TRAVMINS	19.2%	19.1%	18.7%	19.7%	19.4%	18.3%
V46	1.0%	1.8%	0.8%	1.3%	2.6%	0.8%

In rank order the two best performing variables were PRIVREN (the percentage of households who rent from a private landlord or letting agency) and NOKIDSP (the percentage of families with no dependent children). PRIVREN, however, was the worst performing variable on three of the four household/individual level variables and so cannot be considered a clear alternative to POPDENS.

As mentioned earlier, although POPDENS was the worst performing variable of the six considered the DfT are somewhat keen to retain an urban/rural dimension when selecting PSUs. In addition, given that the analysis has not identified a clear alternative third stratifier cautiousness may well dictate that there is nothing to be lost by retaining POPDENS as the third stratifier.

Using the formula shown in Section 2.4 the percentage gain in precision by using the R^2's from two models (one containing the potential final stratifier and one existing stratifier (POPDENS)) was computed. The results are shown in Table 10. (Estimated gains of two or more percentage points are shown in shaded cells. A negative sign indicates a loss of precision compared to POPDENS).

Table 10 **Estimates of change in precision of NTS variables using TWO5KM, TWENTYKM, WRKCAR, NOKIDSP, and PRIVREN as the third stratifier in place of POPDENS**

	Estimated reduction in between-PSU variance				
NTS variable	*NTSREG + CARS0P + TWO5KM*	*NTSREG + CARS0P + TWENTYKM*	*NTSREG + CARS0P + WRKCAR*	*NTSREG + CARS0P + NOKIDSP*	*NTSREG + CARS0P + PRIVREN*
BUS15MIN	-0.1%	-6.8%	-12.6%	-13.2%	-14.2%
BIKEINHH	-0.1%	0.1%	0.1%	0.2%	-0.1%
WALKRAIL	2.5%	0.4%	2.0%	2.3%	4.0%
DRIVLICE	-0.1%	0.6%	0.7%	7.4%	-0.2%
SHOPPING	-2.1%	-2.2%	-2.0%	-1.3%	-0.6%
J29	1.4%	1.1%	0.9%	2.0%	2.4%
JOTXSC	1.7%	1.3%	2.9%	1.6%	5.1%
JD	3.3%	5.3%	-1.3%	3.1%	0.5%
JJXSC	-0.3%	-1.3%	7.0%	-1.9%	2.5%
JDSCHOOL	-2.6%	-1.1%	-3.4%	-2.1%	-3.0%
STGEWALK	1.6%	0.7%	10.6%	1.0%	11.0%
STGECAR	-3.6%	-3.8%	11.3%	-3.7%	9.3%
STTXSC	1.6%	1.3%	2.4%	1.8%	3.2%
SD	2.9%	4.8%	-1.4%	2.5%	-0.2%
S25	2.7%	4.7%	-1.5%	2.4%	-0.5%
S36	1.1%	0.9%	0.4%	1.5%	1.2%
TRAVMINS	1.1%	0.9%	0.5%	1.6%	1.3%
V46	0.2%	1.0%	0.0%	0.5%	1.8%

For seven of the 18 NTS variables no alternative PSU stratifier achieved an estimated gain in precision of two or more percentage points. Only NOKIDSP achieved a gain of two or more percentage points for more than one of BUS15MIN, BIKEINHH, WALKRAIL and DRIVLICE (the household/individual level variables).

Table 10 also clearly shows that the optimal stratifiers for one NTS variable may be sub-optimal for others. For example, WRKCAR produced gains in precision (above two percentage points) for WALKRAIL, JOTXSC, JJXSC, STGEWALK, STGECAR and STTXSC, but losses for BUS15MIN, SHOPPING and JDSCHOOL.

As noted in Section 2.4, the problem of 'over-fitting' means that the results in this table must be viewed with some caution. For this reason the above analysis was repeated using the NTS 2004 for a subset of the outcome measures. Note, however, that due to the `quasi-panel' design described in Section 1.2 half of the 2004 PSUs were carried over from 2003. Hence there still remains an important overlap between the 2003 and 2004 PSU level datasets used for this report. The percentage gain in precision compared to using POPDENS as the third stratifier are shown in Table 11.

Table 11 **Estimates of change in precision of NTS variables using TWO5KM, TWENTYKM, WRKCAR, NOKIDSP, and PRIVREN as the third stratifier in place of POPDENS (using the NTS 2004)**

	Estimated reduction in between-PSU variance				
NTS variable	*NTSREG + CARS0P + TWO5KM*	*NTSREG + CARS0P + TWENTYKM*	*NTSREG + CARS0P + WRKCAR*	*NTSREG + CARS0P + NOKIDSP*	*NTSREG + CARS0P + PRIVREN*
J29	0.6%	0.6%	4.3%	1.6%	2.3%*
JOTXSC	0.3%	-0.4%	5.2%*	0.7%	2.0%*
JD	0.8%	5.0%*	-0.3%	5.9%*	2.2%
JJXSC	0.5%	-0.8%	1.3%	-0.8%	-0.3%
STTXSC	0.5%	-0.4%	4.3%*	0.6%	1.2%
SD	1.2%	6.1%*	-0.4%	6.0%*	1.9%
S25	1.1%	6.1%*	-0.5%	5.9%*	1.7%
S36	0.7%	0.6%	3.3%	1.1%	1.5%
V46	-0.2%	3.7%	-0.2%	-0.7%	-0.5%

Table 11 shows a fairly close agreement between the NTS 2003 and 2004 results. (An asterisk denotes an estimated two percentage point increase in both survey years). Although POPDENS may not be fully optimal in terms of predicting the NTS variables the results did not indicate one variable which performed consistently better than all others.

3.6 Reconsidering the regional classification

Having chosen to retain CARS0P and POPDENS a final step was to test if NTSREG still remained an optimal choice of regional stratifier. Two models were tested:

Model 1: NTSREG × CARS0P (3 bands) + POPDENS
Model 2: GOR × CARS0P (3 bands) + POPDENS

The percentage gain in precision by using GOR in place of NTSREG was computed. The results are shown in Table 12. The results again show that NTSREG performs better than GOR for the four household/individual level variables, confirming the usefulness of NTSREG as a geographical classification.

Table 12 Estimates of change in precision of NTS variables using GOR as the first stratifier in place of NTSREG

NTS variable	Estimated reduction in between-PSU variance
BUS15MIN	-2.8%
BIKEINHH	-1.7%
WALKRAIL	-8.4%
DRIVLICE	-6.1%
SHOPPING	1.0%
J29	1.4%
JOTXSC	1.4%
JD	2.7%
JJXSC	-0.2%
JDSCHOOL	-1.5%
STGEWALK	1.8%
STGECAR	-2.8%
STTXSC	0.7%
SD	1.9%
S25	1.9%
S36	-0.6%
TRAVMINS	-0.7%
V46	2.4%

4 CONCLUSIONS

The current set of postcode sector stratifiers for the NTS consist of a regional classification based on NUTS2 areas and Census measures of car ownership (percentage of households with no car) and population density.

Gains in the precision of survey estimates can be achieved if the stratifiers chosen are highly correlated to the key survey variables. (Such gains, however, affect only the between-areas component of sampling variance). Using the NTS 2003 data (aggregated to postcode sector level) in tandem with 2001 Census information a series of regression models were fitted to examine which possible stratifiers contributed most to predicting key survey dependent variables.

The current NTS regional stratifier performed better than GOR in predicting the household/individual level variables. Where GOR did perform better than NTSREG the difference in adjusted R^2 was less than one percentage point. NTSREG remains, therefore, the optimal choice as first stratifier.

Optimal stratifiers for one variable often turn out to be sub-optimal for others because the correlation between survey variables and stratification factors vary for each survey measure. As a result, the variables chosen are usually those which appear in several optimal models and where there is a strong intuitive reason for their strength as stratifiers (Bruce,1993). Having taken account of the 40 NTS regions, car ownership measures featured in eight of the 18 models, and were the best performing variable in six cases. These were spread across different levels of the hierarchical database (i.e. household, individual, trips and stage).

CARS0P (the percentage of households with no car), therefore, ought to be retained as the second stratifier.

Whilst variables such as WRKCAR and NOKIDSP perform better than POPDENS in predicting a number of NTS variables the DfT have argued that using population density as a stratifier serves an important *general purpose* as it ensures a balanced sample of households living in urban and rural areas. Given its general purpose, and the fact that no variable performed consistently better than all others, POPDENS ought to be retained as the third stratifier *unless* more optimal measures of urbanity become available at the PSU level.

Finally, increasing the number of car ownership bands from three to four within each NTS region would weaken the effect of the third stratifier. All other things being equal, increasing the number of NTSREG × CARS0P bands from 120 to 160 would result in a less finely tuned selection of PSUs (in terms of population density) within each band. For this reason it is recommended to continue to use three car ownership bands within each NTS region.

5 REFERENCES

Barton, J. (1996) *Investigating stratification options for the National Travel Survey.*

Bruce, S. (1993) Selecting stratifiers for the Family Resources Survey. *Survey Methodology Bulletin 32*: 20-25.

Hayllar, O., McDonnell, P., Mottau, C. and Salathiel, D. (2005) *National Travel Survey 2003 & 2004 Technical Report* (National Centre for Social Research: London).

Insalaco, F. (2000) Choosing stratifiers for the General Household Survey. *Survey Methodology Bulletin 46*: 6-14.

Korovessis, C. (2001) Sampling minority ethnic groups in the UK population. *Survey Methods Newsletter 21*: 1 12-19.

Pickering, K., Tipping, S. and Scholes, S. (2005) *Weighting the National Travel Survey: Methodology Final Report*. (http://www.dft.gov.uk/stellent/groups/dft_transstats/documents/page/dft_transstats_041302.pdf)

Thomas, R. and Pickering, K. (2003) *Methodological Review of the Survey of English Housing* (National Centre for Social Research: London).

Wooldridge, J. (2003) *Introductory Econometrics: A Modern Approach, 2nd edition* (Thomson South-Western: United States).