Natural England Commissioned Report NECR268

Small Area Estimation feasibilty: MENE survey

First published 28th August 2019

www.gov.uk/natural-england

Foreword

Natural England commission a range of reports from external contractors to provide evidence and advice to assist us in delivering our duties. The views in this report are those of the authors and do not necessarily represent those of Natural England.

Background

Using data collected by the Monitor of Engagement with the Natural Environment (MENE survey), this project aims to determine whether Small Area Estimation (SAE) techniques can be used by Natural England to improve upon the accuracy and precision of survey based estimates of the extent to which local populations engage with the natural environment

This report should be cited as:

CITY SCIENCE. 2019. *Small Area Estimation feasibility: MENE survey*. Natural England Commissioned Reports, Number268.

Natural England Project Manager - Rose O'Neill, Principal Specialist
Rose.O'Neill@naturalengland.org.uk

Contractor - City Science

Keywords - MENE, Small Area Estimation, engagementment, nature

Further information
This report can be downloaded from the Natural England Access to Evidence Catalogue: **http://publications.naturalengland.org.uk/** . For information on Natural England publications contact the Natural England Enquiry Service on 0300 060 3900 or e-mail **enquiries@naturalengland.org.uk**.

ISBN 978-1-78354-534-6

Natural England

Small Area Estimation feasibility: MENE survey

CITY SCIENCE
endless possibilities

1 Project Aims

1.1 Overview

1. Using data collected by the *Monitor of Engagement with the Natural Environment (MENE survey)*, this project aims to determine whether Small Area Estimation (SAE) techniques can be used by Natural England to improve upon the accuracy and precision of survey-based estimates of the extent to which local populations engage with the natural environment. More specifically, it is concerned with quantifying improvements in accuracy and precision and determining what savings (in terms of effective sample sizes) can be achieved when applying SAE techniques at Local Authority level as well as illustrating the potential for use at the Middle Layer Census Super Output Area (MSOA).

2. Specifically for this evaluation, the request was to model responses to “Q1: Volume of visits per day over last 7 days” and “Q17: Frequency of visits during last 12 months”. Whilst an interesting range of attitudinal and other questions are asked which may be associated with these responses, in order to conduct small area estimation we have to match predictor variables which ones which can be found in some auxiliary data. The variables identified as being of potential interest were:

- Age
- Sex
- Ethnicity
- Marital status
- Working status
- Socio-economic group

- Lifestage
- Household size
- Children in household
- Adults in household
- Tenure

It is also possible to use the geographical information provided.

1.2 *MENE survey data*

3. As currently available to us, the MENE survey comprises 420,790 responses. These have been accumulated over *nine* years of the survey.

Dates	Responses
March 2009 - February 2010	48,514
March 2010 - February 2011	46,099
March 2011 - February 2012	47,418
March 2012 - February 2013	46,749
March 2013 - February 2014	46,785
March 2014 - February 2015	45,225
March 2015 - February 2016	45,965
March 2016 - February 2017	46,558
March 2017 - February 2018	47,477

Table 1.1: Number of responses in the MENE survey data by year of survey

1.3 *Comparison of variables in MENE survey and Census Small Area Microdata*

4. In order to perform small area estimation it is necessary to match relevant predictor variables within the survey data with comparable variables in the auxiliary data. The hope is that by applying statistical evidence from the survey (with limited coverage) to auxiliary data with universal coverage it is poissble to obtain small area estimates based on the survey which have universal coverage. This feasibility study concentrates on using 2011 census Small Area Microdata (SAM) as auxiliary data. These SAM data allows for a full cross-classification of all predictor variables. This can be helpful in the event of interactions between predictor variables. The problem is that there are often differences in the way different variables are classified in different census outputs. For example, age can be banded in very different ways. Social factors (such as occupation types) can be collapsed in different ways. The first task is therefore to review the potential matches between the potential auxiliary variables and survey variables.

1.3.1 Age

5. As noted above, even a clearly conceptualised variable such as age can be banded differently in different census tables. Tables 1.2 and 1.3 record the number of respondents in the MENE survey and the SAM by designated age band. It can be seen that more age bands are available in the SAM data, however there is a direct 1:1 mapping of a collapsed version of the SAM age bands onto the MENE categorization.

16-24	56,032
25-34	69,405
35-44	65,721
45-54	62,452
55-64	57,812
65+	109,368

Table 1.2: MENE Respondents by age band

16 - 18	175,525
19 - 24	158,461
25 - 29	201,523
30 - 34	110,051
35 - 39	255,949
40 - 44	194,748
44 - 49	185,159
50 - 54	187,354
55 - 59	205,389
60 - 64	205,093
65 - 69	179,955
70 - 74	158,954
75 - 79	169,477
80 - 84	133,764
85 - 89	108,696
90+	88,494

Table 1.3: SAM respondents by age band

1.3.2 *Sex*

Tables 1.4 and 1.5 (perhaps obviously) match well in terms of sex. It does look as if the MENE survey response rate might have been *slightly* higher amongst females.

Female	222,849
Male	197,941

Table 1.4: MENE Respondents by sex

Female	1,446,402
Male	1,401,747

Table 1.5: SAM respondents by sex

1.3.3 *Car ownership*

6. Tables 1.6 and 1.7 tabulate the responses to questions on car ownership in the MENE survey and SAM data respectively. These are very likely key predictor variables for access to the environment. The MENE survey data only records whether or not a car is available, the census microdata are more granular. However, it is clearly possible to map the SAM data to the MENE survey and use presence or absence of a car as both a predictor variable in modelling the survey response as well as an auxiliary variable.

No	98,871
Yes	232,270
NA.	89,649

Table 1.6: MENE Respondents by car ownership

0	539,626
1	1,084,978
2	859,943
3	221,962
4 or more	88,511

Table 1.7: SAM respondents by car ownership

7. However, an obvious problem with the car ownership variable appears to be in relation to the years when this variable was recorded. Table 1.8 illustrates the problem. Car ownership stopped being recorded after survey year 2015/2016.

	No	Yes	NA
Y0910	13,768	34,746	0
Y1011	12,893	33,206	0
Y1112	13,915	33,503	0
Y1213	14,668	32,081	0
Y1314	14,105	32,680	0
Y1415	14,346	30,879	0
Y1516	13,913	32,052	0
Y1617	1,263	3,123	42,172
Y1718	0	0	47,477

Table 1.8: MENE Respondents by car ownership by year

1.3.4 *Self reported health*

8. Tables 1.9 and 1.10 tabulate the responses to the census and MENE survey in relation to self reported general health. These variables appear to match up well in terms of definition, however, there appears to be a large number of non-responses in the MENE data which limits the potentially use of this variable for small area estimation.

Don t know	41
Very bad	503
Bad	2,396
Fair	8,482
Good	18,592
Very good	11,830
NA	378,946

Table 1.9: MENE Respondents by health status

Bad	121,091
Fair	369,581
Good	956,602
Very bad	35,840
Very good	1,330,176
NA	34,859

Table 1.10: SAM respondents by health status

9. Table 1.11 breaks down the MENE responses by year, showing firstly, that this question only started to be asked in 2014/2015 and secondly that only around 10% of respondents had a response recorded.

	Don't know	Very bad	Bad	Fair	Good	Very good	NA
Y0910	0	0	0	0	0	0	48,514
Y1011	0	0	0	0	0	0	46,099
Y1112	0	0	0	0	0	0	47,418
Y1213	0	0	0	0	0	0	46,749
Y1314	0	0	0	0	0	0	46,785
Y1415	9	84	535	1,915	4,286	2,778	35,618
Y1516	10	132	577	2,203	4,769	2,985	35,289
Y1617	13	127	616	2,144	4,813	3,002	35,843
Y1718	9	160	668	2,220	4,724	3,065	36,631

Table 1.11: MENE Respondents by general health by year

1.3.5 *Working status*

10. Tables 1.12 and 1.13 record some information on working status. It would appear that it might be possible to collapse both variables to some form of Student, Unemployed, Part Time Work, Full Time Work, Retired if that is necessary. Figure 1.14 provides further details on the working hours for those census respondents where this is relevant.

It might be possible to provide a mapping for the MENE survey data, but it is not clear that a perfect mapping is possible and given the time limitations on this feasibility study it will not be considered further here.

Full time	144,933
In Education	25,967
Not working	73,234
Part time	52,774
Retired	123,882

Table 1.12: MENE Respondents by working hours

Employee F Economical	795,670
Employee P Economical	285,961
Full time Economical	71,437
Looking af Economical	90,131
Other Economical	49,482
Permanentl Economical	88,703
Retired Economical	485,756
Seeking or Economical	89,846
Self emplo Economical	201,934
Student Economical	124,287
NA	564,942

Table 1.13: SAM respondents by economic activity

Full time 31 to 48 hours	769,355
Full time 49 or more hours	177,317
Part time 15 hours or less	129,589
Part time 16 to 30 hours	261,719
NA	1,510,169

Table 1.14: SAM respondents by working hours

1.3.6 Socio-economic grouping

11. Tables 1.15 and 1.16 shows that the MENE survey and SAM data match in usage of the shorter version of Socio-economic groupings. The MENE survey does have a fuller listing (shown in table 1.17) but there is insufficient information in the SAM to match.

AB	76,756
C1	111,873
C2	85,766
DE	146,395

Table 1.15: MENE Respondents by socio-economic group

AB	404,351
C1	539,826
C2	393,856
DE	444,695
NA.	1,065,421

Table 1.16: SAM respondents by socio-economic group

A	12,123
B	64,633
C1	111,873
C2	85,766
D	67,074
E	79,321

Table 1.17: MENE Respondents by socio-economic group (full listing)

1.3.7 *Tenure*

12. The MENE survey uses a shorter form of tenure classification (table 1.18) which does not match exactly the shorter form of the SAM tenure classification (table 1.19). However, there is a longer version of the tenure classification in the SAM (table 1.20) and it may be possible to get a reasonable mapping. For the purposes of this feasibility study, this will not be pursued further. For example, there is a need to determine whether the SAM "Rented from registered social landlord or housing association" maps onto the MENE "Rent local authority" or the MENE "Rent Private".

Mortgage	102,790
Other	17,339
Owned outright	137,973
Rent local authority	78,197
Rent private	84,491

Table 1.18: MENE Respondents by tenure

Owns outright	717,054
Owns with mortgage	1,103,200
Rent free	30,594
Rents	925,056
Shared ownership	19,116
NA	53,129

Table 1.19: SAM respondents by tenure

Employer of household member	8,112
Other	7,819
Owns outright	717,598
Owns with mortgage	1,103,584
Private landlord or letting agency	427,365
Relative or friend of household member	24,177
Rent free	30,395
Rented council	247,163
Rented registered social landlord or housing assoc	207,239
Shared ownership	19,177
NA	55,520

Table 1.20: SAM respondents by tenure (tenduk11)

1.3.8 *Ethnicity*

13. Tables 1.21 and 1.22 denote responses on the census as tabulated in the SAM and the MENE survey in relation to ethnicity. These are possibly not identical, but it appears that it would be possible to match on a collapsed, simplified classification.

Table 1.21: MENE Respondents by ethnicity

African	9,749
Any other	1,791
Any other Asian backgrund	6,067
Any other Black background	1,055
Any other mixed background	1,286
Any other white background	27,812
Bangladeshi	3,831
Caribbean	6,601
Chinese	1,738
Indian	11,678
Pakistani	8,703
Refused	2,008
White & Asian	925
White & Black African	787
White & Black Carib bean	1,800
White British	331,276
White Irish	3,683

Table 1.22: SAM respondents by ethnicity

Asian Bangladeshi	22,297
Asian Chinese	20,825
Asian Indian	71,919
Asian Other	42,750
Asian Pakistani	56,670
Black African	49,745
Black Caribbean or Other	43,846
Mixed White and Asian or Other	31,670
Mixed White and Black	29,556
Other Other	28,729
White Irish	26,741
White Other	131,211
White UK	2,257,331
NA	34,859

14. Consequently, the MENE and SAM variables will be mapped onto a standard five level ethnic classification; “White British”, “White Other”, “Black”, “Asian”, “Other” as shown in table 1.23

asian	32,017
black	17,405
other	8,597
white_british	331,276
white_other	31,495

Table 1.23: MENE Respondents by ethnicity, mapped to five categories

1.3.9 Marital Status

15. Tables 1.24 and 1.25 depict the classification of marital status by the two surveys. There is no perfect match, but it appears possible to apply a simpler mapping.

Married	238,230
NA	41
Sep/Wid/div	73,469
Single	109,050

Table 1.24: MENE Respondents by marital status

Divorced or from legally dissolved civil partnership	204,529
Married	1,060,915
Registered civil partnership	5,394
Separated	59,938
Single (never married/never partnership)	1,358,769
Widowed or Surviving partner of civil partnership	158,604

Table 1.25: SAM respondents by marital status

1.3.10 *Disability*

16. Tables 1.26 and 1.27 depict the classification of self reported disability status by MENE survey and SAM. These could be difficult to reconcile, as we have no way of understanding how a person who entered some limitations on the census would have recorded a binary indicator on the MENE survey. Specifically, it may be that some people who on the census regard themselves as “limited a little” would not indicate “Yes” to a question on disability in the MENE survey. This is a known problem when attemption to reconcile the 2001 census (which also has a yes/no classification for limiting long term illness) with the 2011 census. For the feasibility study, this variable will be ignored. If, as seems likely, it is an important predictor of engagement with the natural environment it could be included in future use, but this would require a degree of validation as to how best to map “disabled / not disabled” MENE responses onto “limited a little”, “limited a lot”, “not limited” SAM responses. This could potentially be possibly using a latent variable construnction (where we posit an underlying latent or hidden variable which “measures” disability level on a continuous numeric scale and attempt to estimate cut points which identify above and below which someone on MENE identifies as having a disability and someone on SAM reports themself as having no, little or very limiting long term illness. The conclusions will recommend that there could be great value in ensuring in future that MENE survey questions align closely with definitions found in such auxiliary data as the census.

No	260,144
Yes	70,997
NA.	89,649

Table 1.26: MENE Respondents by self-reported disability status

Limited a little	263,047
Limited a lot	238,435
Not limited	2,311,808
NA	34,859

Table 1.27: SAM respondents by self-reported disability status

17. However, a bigger problem with the use of this variable is that the question has not been asked in all years. Table 1.29 shows that the question was only asked until 2015/16.

18. However, a bigger problem with the use of this variable is that the question has not been asked in all years. Table 1.29 shows that the question was only asked until 2015/16.

	No	Yes	NA
Y0910	38,220	10,294	0
Y1011	36,674	9,425	0
Y1112	37,421	9,997	0
Y1213	36,874	9,875	0
Y1314	36,867	9,918	0
Y1415	35,304	9,921	0
Y1516	35,427	10,538	0
Y1617	3,357	1,029	42,172
Y1718	0	0	47,477

Table 1.28: MENE Respondents by disability status by year

	No	Yes	NA
Y0910	38,220	10,294	0
Y1011	36,674	9,425	0
Y1112	37,421	9,997	0
Y1213	36,874	9,875	0
Y1314	36,867	9,918	0
Y1415	35,304	9,921	0
Y1516	35,427	10,538	0
Y1617	3,357	1,029	42,172
Y1718	0	0	47,477

Table 1.29: MENE Respondents by disability status by year

1.3.11 *Family structure*

19. Tables 1.30 and 1.31 are comparable in terms of indicating the number of members of a household up until 4. The MENE then records all households of 5 and more together whereas the SAM breaks down 5 person, 6 person and 7 and more person households. Clearly, it is possible to map the SAM data to the MENE classification. Further variables record number of children and family structure. It may be possible to provide some reduced common variable on family structure if that is important.

1	91,416
2	141,593
3	74,636
4	69,202
5+	43,943

Table 1.30: MENE Respondents by number in household

1	358,268
2	811,983
3	558,114
4	613,164
5	273,641
6	116,467
7 or more	60,913

Table 1.31: SAM respondents by numbers in household

20. There are further variables in the MENE dataset which break down households by the numbers of children and adults in a household (figures 1.32 and 1.33). It is perhaps possible to compare the first table (number of children) with data given by the census shown in table 1.34. The census further more gives information on a respondents status within a househol. However, such mappings are not trivial and will not be pursued for this feasibility study.

Any	123,588
None	297,202

Table 1.32: MENE Respondents by number of children in household

1	109,755
2	208,801
3	58,534
4	30,138
5+	13,562

Table 1.33: MENE Respondents by number of adults in household

None	1,010,346
One aged 0 9	218,407
One aged 10 18	228,977
Three or more youngest aged 0 9	244,047
Three or more youngest aged 10 18	42,000
Two youngest aged 0 9	331,787
Two youngest aged 10 18	163,605
NA	608,980

Table 1.34: SAM respondents by number of children in household

1.3.12 *Lifestage*

21. The MENE survey contains a demographic variable “lifestage”. There is no simple mapping onto any variable in the census SAM release, although it may be possible to construct some kind of common variable.

1.3.13 *Areal level variables*

22. Further variables are available in the MENE survey describing the area rather than the individual. It is our view that given we are provided with the respondent postcode sector, it is possible to obtain a much wider range of geospatial information than that already contained in the MENE survey. This would also, by definition, be immediately available as an auxiliary variable. It would also be possible to incorporate this in a multilevel model, where we could allow for the fact that person-types in a similar area might have some commonality in their tendency to engage with the natural environment. More importantly, with a little more development time, it would be possible to develop these in a fully spatial model, and allow for a spatial correlation in the way people engage with the environment as well as explicitly using geographical factor, such as ease of access to green space to inform a modelling exercise. These are not considered here further due to the time restrictions on the feasibility study, but in our view would provide powerful additional insight in their own right, as well as strengthening the small area estimation.

1.3.14 Missing data

23A dominant feature of this exploration is that there are a number of systematically missing responses. Table 1.35 lists the variables that will be considered for this feasibility study. The inclusion of Car Ownerhip means that, for modelling purposes, we will only consider data up until the survey year 2015/2016.

Age	All years
Sex	All years
Car Ownership	Until 2015/16
Socio-economic grouping	All years
Ethnicity	All years (Map to 5 groups)
Marital status	All years (map to MENE)
Children in household	All years (create binary indicator)
Number in household	All years (map to MENE)

Table 1.35: Summary of variables considered for modelling purposes

24In Bayesian models, it is relatively straightforward to include a data imputation stage to handle missing data. This can easily deal with small numbers of randomly missing responses from individual responses. It cannot however be assumed that it magically corrects two years of systematic non-response. As this study is intended as a quick feasibility study, the data will be subsetted to include data from survey year 2009/2010 through to 2015/2016 for variable screening and local small area prediction. For overall sample size work, it will use just data from 2015/2016.

1.3.15 Response variables

25Finally, table 1.36 and table 1.37 summarise the responses of interest for Question 1 and Question 17 respectively. For Q1 the responses have been dichotomised to indicate whether a respondent stated they had taken a trip to the natural environment in the last week. Question 17 has been dichotomised to indicate whether in the past year the respondent reported having taken a trip at least once a month.

26Most apparent from table 1.37 is that only around a quarter to a fifth of respondents answer the question on trips in the previous year. Clearly, this reduces the sample size for the analysis considerably for this response variable. However, a modelling approach can help with this if it is reasonable to use data from more than one year at a time assuming any underlying trends can be modelled accurately.

Table 1.36: MENE trips in last week

	No	Yes	NA
Y0910	28,140	20,374	0
Y1011	28,710	17,389	0
Y1112	28,404	19,014	0
Y1213	28,564	18,185	0
Y1314	27,977	18,808	0
Y1415	26,567	18,658	0
Y1516	27,536	18,429	0
Y1617	25,958	20,600	0
Y1718	24,471	23,006	0

Table 1.37: MENE at least one trip a month in previous twelve months by year

	No	Yes	NA
Y0910	3,121	7,986	37,407
Y1011	2,889	7,741	35,469
Y1112	2,742	7,845	36,831
Y1213	2,864	7,680	36,205
Y1314	2,642	7,910	36,233
Y1415	2,554	7,917	34,754
Y1516	2,680	7,996	35,289
Y1617	2,180	8,535	35,843
Y1718	2,186	8,660	36,631

1.4 Exploratory Modelling

27. Exploratory modelling has been carried out using the R software[1] version 3.5.3 "Great Truth".

28. The AIC (*A*n *I*nformation *C*riterion[2]) is an overall metric which assesses comparative model fit, models with a lower AIC are better fitting than those with a higher AIC. It contains a fit term and a penalty term. It is always possible to improve model fit by adding more variables, the penalty term acts as a deterrent against including too many variables. As implemented here, the automated model search starts with a full model (all possible terms) and checks sequentially whether removal of any one variable would notably improve the AIC, and if so which variable is the best to remove. The final model is one which has as few variables as possible, but retains comparable predictive power to the full model.

[1] R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[2] Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", in Petrov, B. N.; Csaki, F. (eds.), 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akademiai Kiado, pp. 267-281. Republished in Kotz, S.; Johnson, N. L., eds. (1992), Breakthroughs in Statistics, I, Springer-Verlag, pp. 610-624.

29. For the purposes of the feasibility study, given the complexity of dealing with seven years of survey data[3] a quick variable selection method based on AIC was applied. Specifically, the `stepAIC()` function from the MASS package[4]. This choice was made on pragmatic grounds; in order to identify candidate models within the scope the feasibility study. The aim is to select a model which can provide an adequate representation of the survey data and which is sufficient to give a realistic prospect for a small area estimation process to be accurate. Using the AIC provides a reasonable method for variable selection which can find a candidate model without overfitting to the specific dataset examined.

[3] As noted in the previous section, the most recent two years of data were disregarded as key questions on car ownership had not been asked

[4] Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

1.4.1 Model fitting to two dichotomised outcomes

30. The response to MENE survey Question 1 records the number of trips taken in the past seven days. Question 17 has been dichotomised as at least one trip a month. A much smaller number of respondents reply to Q17 than do to Q1. For the purposes of this feasibility study, we have modelled both as a binary indicator, equal to 0 if either no trips in the last seven days, or a trip frequency of less than once a month in the previous year. The outcome was denoted as 1 if at least one trip had been made in the last seven days or the trip frequency was at least one trip a month in the previous year. A response "1" therefore indicates the positive outcome and all parameter estimates can be interpreted in that regard.

31. All the available predictors were examined, including two way interactions. Parameter estimates from the finally selected model are given in table 1.1 and a visual summary is subsequently given in figure 1.2 (along with 95% and 90% confidence intervals for the parameter estimates).

Parameter estimates for predicting Q1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.974	0.027	-36.341	0.000
age25-34	0.064	0.020	3.242	0.001
age35-44	0.002	0.020	0.112	0.910
age45-54	-0.112	0.021	-5.428	0.000
age55-64	-0.103	0.022	-4.714	0.000
age65+	-0.524	0.021	-24.742	0.000
sexMale	0.138	0.020	6.849	0.000
carYes	0.444	0.009	46.953	0.000
segC1	-0.300	0.011	-27.218	0.000
segC2	-0.513	0.012	-43.895	0.000
segDE	-0.720	0.011	-64.582	0.000
marstatMarried	0.060	0.013	4.643	0.000
marstatSep/Wid/div	-0.027	0.014	-1.901	0.057
ethnicity_5black	0.144	0.025	5.741	0.000
ethnicity_5other	0.551	0.031	17.971	0.000
ethnicity_5white_british	0.898	0.016	56.665	0.000
ethnicity_5white_other	0.720	0.020	35.405	0.000
child_in_hhAny	0.220	0.010	21.966	0.000
adults_in_hh2	0.014	0.013	1.076	0.282
adults_in_hh3	-0.086	0.015	-5.820	0.000
adults_in_hh4	-0.072	0.018	-4.002	0.000
adults_in_hh5+	-0.114	0.025	-4.634	0.000
survey_yearY1011	-0.193	0.014	-14.054	0.000
survey_yearY1112	-0.072	0.014	-5.281	0.000
survey_yearY1213	-0.101	0.014	-7.364	0.000
survey_yearY1314	-0.051	0.014	-3.720	0.000
survey_yearY1415	0.020	0.014	1.435	0.151
survey_yearY1516	-0.036	0.014	-2.648	0.008
age25-34:sexMale	-0.257	0.027	-9.501	0.000
age35-44:sexMale	-0.191	0.027	-7.002	0.000
age45-54:sexMale	-0.098	0.028	-3.540	0.000
age55-64:sexMale	-0.123	0.028	-4.401	0.000
age65+:sexMale	0.022	0.025	0.858	0.391

In total, 326727 data points were used to fit this model.

Parameter estimates for predicting Q17

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.851	0.062	13.772	0.000
age25-34	-0.172	0.050	-3.446	0.001
age35-44	-0.405	0.051	-8.023	0.000
age45-54	-0.670	0.050	-13.304	0.000
age55-64	-0.787	0.053	-14.889	0.000
age65+	-1.190	0.050	-23.818	0.000
sexMale	0.225	0.051	4.374	0.000
carYes	0.579	0.021	28.086	0.000
segC1	-0.321	0.030	-10.585	0.000
segC2	-0.599	0.031	-19.317	0.000
segDE	-0.840	0.029	-29.295	0.000
marstatMarried	0.153	0.030	5.112	0.000
marstatSep/Wid/div	-0.084	0.032	-2.656	0.008
ethnicity_5black	0.097	0.049	1.994	0.046
ethnicity_5other	0.457	0.067	6.822	0.000
ethnicity_5white_british	0.860	0.032	26.884	0.000
ethnicity_5white_other	0.641	0.044	14.706	0.000
child_in_hhAny	0.248	0.025	9.819	0.000
adults_in_hh2	-0.034	0.029	-1.148	0.251
adults_in_hh3	-0.175	0.034	-5.121	0.000
adults_in_hh4	-0.109	0.043	-2.571	0.010
adults_in_hh5+	-0.173	0.055	-3.160	0.002
survey_yearY1011	0.054	0.032	1.693	0.090
survey_yearY1112	0.132	0.032	4.109	0.000
survey_yearY1213	0.081	0.032	2.532	0.011
survey_yearY1314	0.211	0.032	6.522	0.000
survey_yearY1415	0.269	0.033	8.265	0.000
survey_yearY1516	0.215	0.032	6.678	0.000
age25-34:sexMale	-0.342	0.069	-4.954	0.000
age35-44:sexMale	-0.284	0.070	-4.055	0.000
age45-54:sexMale	-0.164	0.069	-2.375	0.018
age55-64:sexMale	-0.202	0.069	-2.947	0.003
age65+:sexMale	-0.066	0.061	-1.091	0.275

In total, 74559 data points were used to fit this model.

Figure 1.1: Parameter estimates from fitting model to Q1 (first four columns) and Q17 (last four columns)

32. To illustrate what these estimates are saying, the following should be noted:

- For every categorical variable considered, the parameter estimates indicate "log odds" relative to the reference case.

  log-odds An appendix is available which explains the reason for using a model which gives the effects of different variables in units of "log-odds". A key point to note is that any parameter estimate *below zero* indicates that those characteristics are associated with a decrease in the chances that such an individual reported a trip relative to the reference case. Conversely, with parameter estimates above zero they are more likely to have reported a trip.

  reference case The "reference case" is a level of the variable which can be seen in the previous section but not given in the table of model parameter estimates 1.1. All other estimates are to be understood as the effect relative to this reference case. For example, as this table gives parameter estimates for "Male", the parameter estimates provided are comparing the difference between respondents who select "Male" with the reference response "Female".

- The reference case then is therefore Female, 16-24 year old, in social group AB, without a car, single, Asian, living with no other adults in the household and having no children in the household. The log odds that such a person would have reported a trip in the last seven days are given as -0.974 These values are taken from the first row in table 1.1.

- The probability that this person (she) would be have reported a trip in the last week can be computed as $\frac{\exp(-0.974)}{1+\exp(-0.974)}$ or about 0.274. Full details on the origin of this equation are appended.

- Likewise, the odds that they reported taking at least one trip a month over the last year are estimated as 0.851, and hence the probability that would be have reported at least one trip a month in the last year was $\frac{\exp(0.851)}{1+\exp(0.851)}$ or about 0.701.

33. Hence, for example, when a parameter estimate is indicated for males of 0.1378, this indicates that the log-odds ratio for males indicating they had made a trip in the last seven days was 0.1378 the value of that for females. This means that their odd ratio was 1.1478 times those of females, indicating that, for all other variables held constant, males were more likely to have reported a trip.

34. The effects of different predictors are readily seen in figure 1.2. This shows the log odds ratio of having made a trip in the last seven days or one trip a month in the last year increased where there was access to a car relative to not having a car or when the respondent was male. Trip taking progressively decreased as the social class went from group C1 to C2 to DE relative to the reference group AB. Respondents who indicated that they had children in a household were more likely to have reported a trip for either question, but three or more adults in a household were consistent with being less likely to report a trip. Figure 1.2 also gives a visual summary of the confidence intervals for each parameter estimate, and compares the models fitted to Q1 and Q17. It can be seen for example that the interval estimates are wider when fitting a model to Q17, reflecting greater uncertainty in the parameter estimate as a result of the smaller sample size. In an informal sense, this plot also captures a sense of "statistical significance". Interval estimates which do not overlap zero would conventionally be regarded as "statistically significant". However, the further the estimate is away from zero, the stronger the relationship between that level of the explanatory variable and the response variable of interest (Q1 or Q17). The width of the interval estimate also visualises the precision or otherwise by which the parameter has been estimated, with wide lines indicating that there is relatively less evidence in the survey sample to estimate the relationship with precision.

35. There does appear to be an uneven trend over time, with the odds for having reported a trip increasing over the seven years of the survey. However, this is largely regarded as a nuisance parameter. In this analysis, the trend over time is not of interest in its own right. Time is included to allow strength to be borrowed from the data over all years without misrepresenting current respondent replies.

36. The parameter estimates for age and sex a slightly more complex as they "interact". In other words, the "effect" of being male is slightly different at different ages. This is perhaps best illustrated visually in figure 1.3. The interaction effects are modest, and mainly appear to relate to either end of the age range. Young males are far more likely to have reported a trip in the last week than would have been the case without an interaction, and old females are less likely to have reported a trip without an interaction effect..

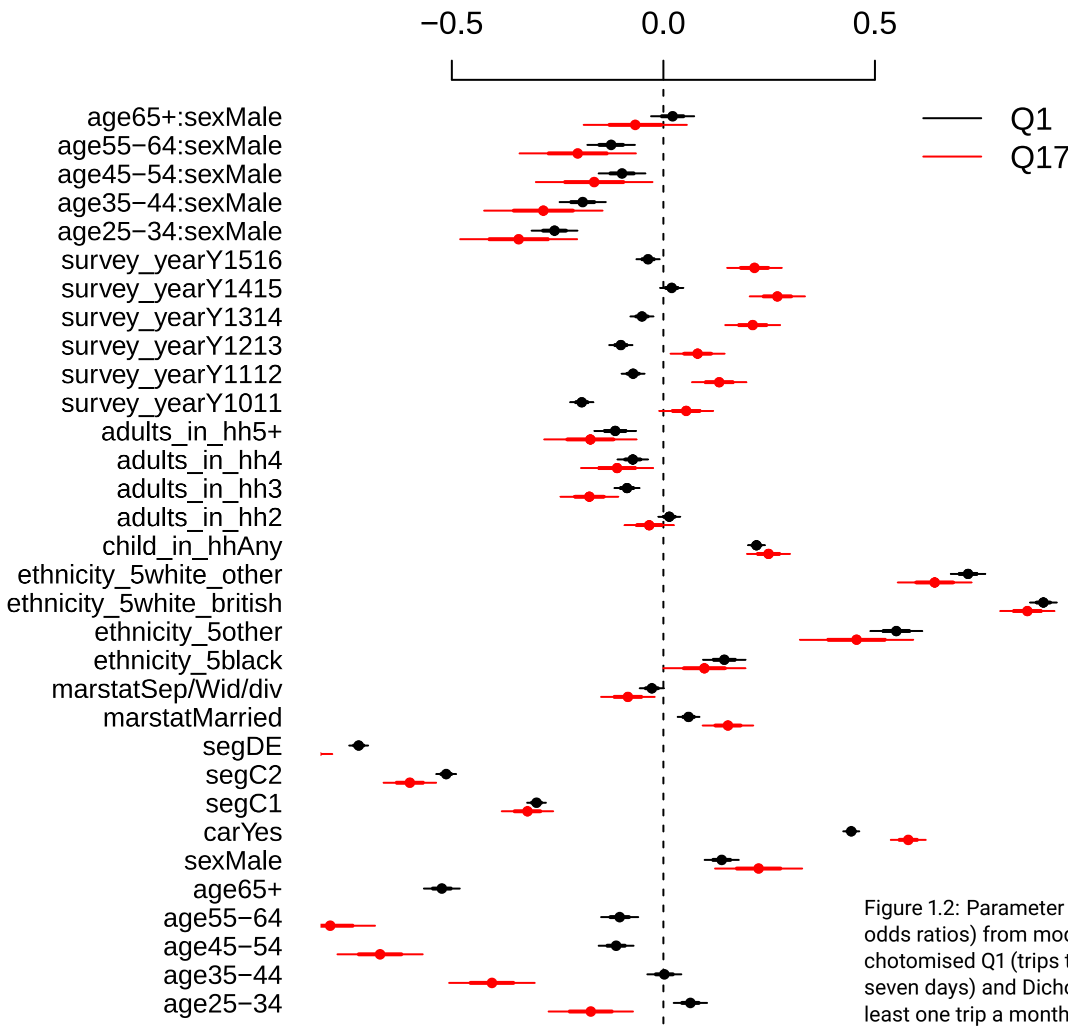

Figure 1.2: Parameter estimates (log odds ratios) from model fitted to Dichotomised Q1 (trips taken in the last seven days) and Dichotomised Q17 (at least one trip a month in the last year)

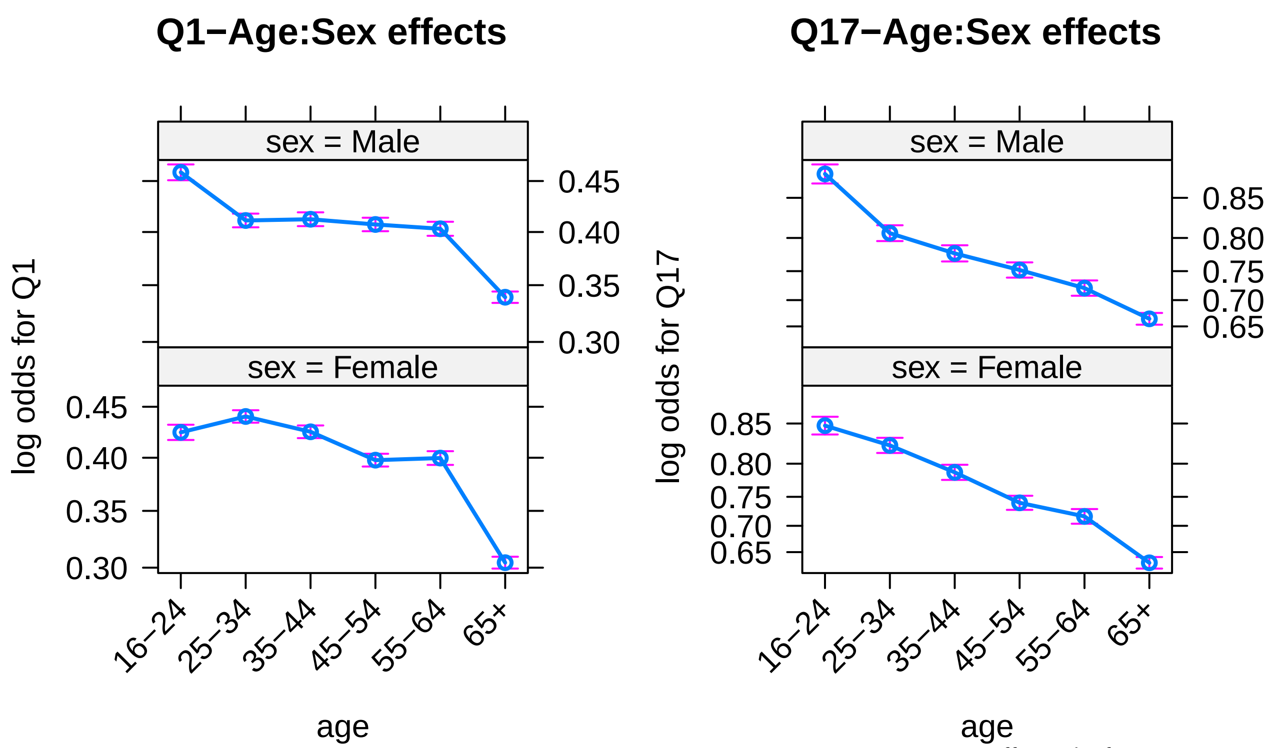

Figure 1.3: Effects plot for age:sex interaction for Q1 and Q17

37. This brief summary therefore represents the model that will be taken forward and used for small area estimation in the next section. Clearly it would be possible to include Bayesian imputation methods for the missing responses to cars availability and use more recent data as well, however this will introduce uncertainty which could impact the precision of the small area estimates.

2 Summary of Methodology

2.1 Small area estimation

38. Small Area Estimation (SAE) techniques have gained increasing acceptance over the last 3 decades[1]. There are a variety of well established approaches, the ones used here draw from model-based survey analysis and seek to "pool" evidence drawn from across a wider sample in order to "enhance" local estimates[2]. They are considered to have some similarities with a post-stratification based approach in design based survey methods.

39. The core idea of this approach to small area estimation is set out in figure 2.1.

There are similarities to post-stratification (a very well established design based approach to survey estimation)[3]. The advantages of a modelling based approach (Bayesian or otherwise) are the way additional multilevel data can be structured. Auxiliary data can be taken from a variety of sources including the relative spatial structure of the sampling points. An advantage of Bayesian methods are that provided the variables used for conventional weighting are used in fitting the models, this accounts for the sampling pattern.

40. Focusing on the behaviour of individual respondents nested within local authorities – a classic "multilevel" scenario – this study has adopted a conventional approach to dealing with hierarchically structured data with a simple count response; namely generalized linear mixed-effects regression modelling. It is an approach that aims to capture, as far as possible, the relationship between the response (or dependent variable) and a number of independent variables which include categorical variables describing individuals' tenure, age, sex, ethnicity, disability, marital status, work status, social classification, lifestage, adults / children in household, size of household, working status, car ownership and their self reported general health[4] The mul-

[1] Rao, J.N.K. (2003) Small Area Estimation New York, Wiley

[2] Gelman, A. and Hill, J. (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models Cambridge, CUP

[3] Andrew Gelman and Thomas C. Little (1997), "Post-stratification Into Many Categories Using Hierarchical Logistic Regression", Technical report, University of Columbia

[4] In order to achieve some parsimony, different models draw upon different sets of categorical variables, for example age may be split into different bands.

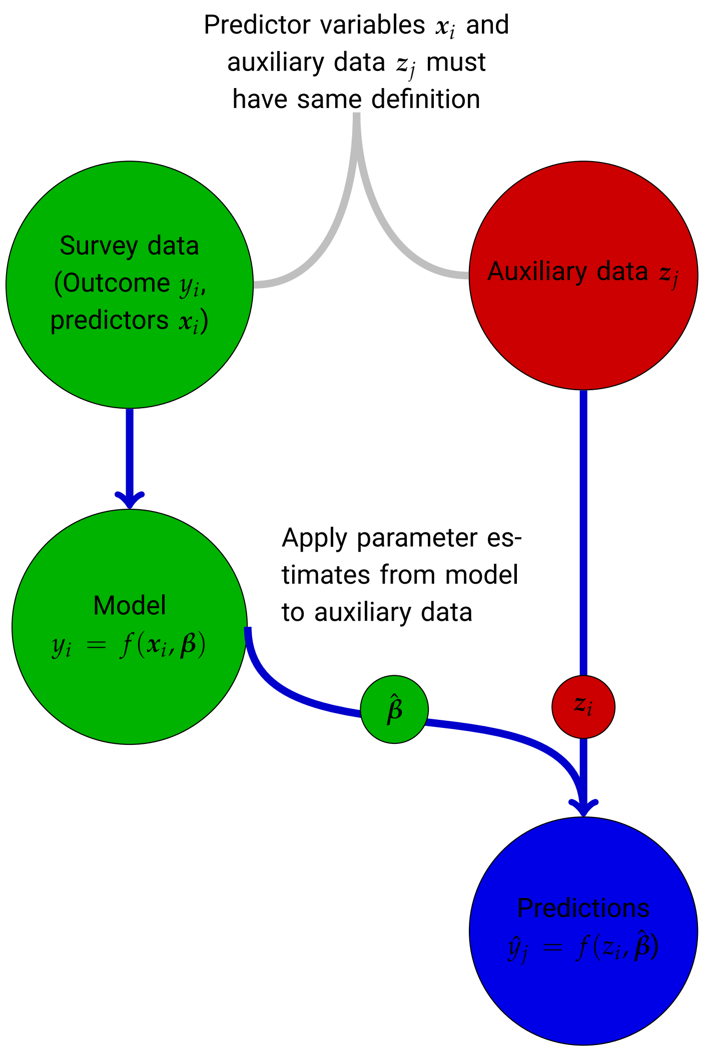

Figure 2.1: Basic outline of synthetic small area prediction

tilevel models are thus used to estimate individual level parameters that explain the relationship between these categorical variables and the response, whilst simultaneously allowing for a random intercept that is unique for each local authority and which describes any systematic authority-by-authority variation in the response that is not explained by the individual-level model.

41. Thus for each individual i in area j, the outcome reports the number of trips made to the natural environment in either the last week or the last year We model this as a Bernoulli distributed random variable so that:

$$Y_{ij} \sim Bernoulli(p_{ij}) \tag{2.1}$$

where the propensity that individual i in area j has reported a given number of trips is given by:

$$logit(p_{ij}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \zeta_j \tag{2.2}$$

where $\boldsymbol{\beta}$ denotes the model parameters and $\boldsymbol{x}$ denote individual characteristics recorded in the survey (such as age, gender, etc.) classified as categorical variables and (in the Bayesian models) ζ denotes the local authority specific random effects.

42. Throughout this report we have assumed that the intercept term varies by local authority, but with the constraint that all these random intercepts are drawn from a common Normal distribution with zero mean and unknown variance. We further extend this model to allow the local authority random intercepts themselves to have a linear predictor. This could be used to capture contextual information that varies by residence rather than by individual. If this were to be undertaken the intercept for Local Authority j would be given as: $z_j^T \boldsymbol{\zeta}$ where ζ denote the upper-level parameters and z denote area-level characteristics which describe various aspects of the local authority as a whole. The linear predictors could be further nested to include, for instance, a term for smaller areas (such as MSOAs) within the local authority.

43. Conventional (i.e. frequentist) approaches were employed for model exploration, but for the final model fit (used to generate the Local Authority estimates) a Bayesian approach was adopted. At this stage each of the “fixed effects” (the parameters relating various factors to the response) were drawn from a Normal distribution with a prior mean of zero and diffuse variance of 0.001. The variance term for the random intercept was assumed to be Uniform (0, 100). These models have been fitted in R using the STAN software package[5].

[5] Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.18.2. http://mc-stan.org/.

44. Having obtained estimates of the individual level parameters, it is possible to apply these to the "known" population structure of a given local authority to produce an estimate of the statistic of interest. This would, in effect, represent the expected response for that local authority given its specific socio-demographic composition. However, the strength of multilevel modelling is that, by estimating individual level effects for typical respondents of each type, the data is used to estimate how responses in each local authority differ from the expected national response. The resulting local authority "random effects" are then used, along with the individual level responses, to make "small area predictions" of the statistic of interest in each local authority.

45. By taking a modelling approach it is also possible to eliminate "nuisance parameters". For example, the trend over time for the years that the survey has been running is obviously of interest in its own right, but is not relevant to attempting to make small area predictions in the same time frame as the most recent data. Consequently, time can be regarded as a nuisance variable, and conditioned out of the model so that all years data can be used to estimate the relationship between predictors and the number of trips made.

46. The Small Area Estimation approach requires that the number of people in each LA with each unique combination of characteristics used in the model is known, or can be estimated. Thus for the engagement models used it is necessary to know how many people are aged 16-24, own a car, are employed, owner occupiers and have a professional occupation – and how many people there are with every other possible combination of characteristics used in the model . This data is obtained by taking the detailed socio-economic composition of Local Authorities from the 2011 Census Small Area Microdata (SAM) 5% Sample, and constraining that distribution to mid-2016 age-sex population totals.

2.2 *Precision and accuracy*

47. Our concern with precision refers to the level of certainty which the various methods attach to their estimates of self-reported engagement with the natural environment. These are conventionally understood with reference to the 95% Confidence Intervals (CIs) associated with the various estimates. More precise measures will have narrower 95% CIs, and any improvement in precision can be measured by comparing the width of SAE-based modelled estimate 95% CIs with the original sample-based direct estimate 95% CIs. Our specific goal here is to express the impact of SAE-based modelling in terms of 'effective sample sizes' and thus predict how much smaller the *MENE survey* samples would need to be in order to achieve the precision achieved by design based survey methods.

48. The measurement of accuracy is less straightforward in that we do not know with any certainty the actual number of people engaging with the natural environment in each Local Authority. There are thus no definitive “correct” values against which to compare the direct and modelled estimates produced using *MENE survey* data.

2.3 *Engagement and LA-Level Estimates*

49. Local Authority random effects, are obviously to be included in the model used to estimate rates of physical activity. These parameters, one for each local authority covered by the analysis effectively measure the extent to which responses in each area differ from what might be expected given the individual level (fixed effects) logistic regression model. These LA-specific intercepts effectively describe how much up or down the fixed effects logistic regression curve must move to best fit the observed data for each local authority.

50. Few of the LA-level fixed effects would be regarded as statistically significant at the 95% level in a conventional analysis. However, this is a mechanism for capturing uncertainty in the modelled responses, and has the effect of reducing the level of uncertainty and hence increasing the precision of many estimates. It would be possible to use better models, such as allowing the random effects to be spatially correlated. This may in turn increase the precision of the subsequent small area estimates.

51. Having obtained “fixed” and local authority “random effect” parameter estimates, these can applied to socio-demographic data for Local Authorities and subsequently, by IPF, to MSOAs. These, as discussed above, are derived from 2011 Census Small Area Microdata, constrained to mid-2016 ONS age-sex totals in local authorities and MSOAs. Thus having determined how many people in each local authority are in each unique 'person-type' cell defined by the model, and having multiplied those sub-populations by the likelihood that those person-types will take a trip to the natural environment, the resulting counts are summed to obtain estimates of the overall number of people in each local authority who took a trip.

52. As already intimated, by modelling the relationship between a range of independent covariates and the dependent variable (in this case whether or not individuals have made a trip in the last week/year) allowance is automatically made for any unintended and unknown bias in the sample. For instance, it is very unlikely that whether or not an individual has a self reported health status of “bad” will ever be used as either a sampling or weighting factor. Yet it is quite likely that among a large number of relatively small sub-samples (which, in this case, are the local authority subsets of the MENE survey there will be a number instances in which a particular sample is very unrepresentative of its underlying population. Given that general health has, as illustrated in , a strong (and of course entirely expected) impact on engagement with the environment, this inevitably means that at least some of the sub-samples will return biased sample-based direct estimates of local levels of physical activity.

2.3.1 *Q1 (trips in the last week): fitting a model to the entire data set*

53. The first set of plots summarise the results of fitting a plot to the entire data set from survey year 2009/2010 through to 2015/2016. In other words, this uses data from all survey years where a question was asked about car availability. Figure 2.2 is a visual summary of the parameter estimates (in terms of log-odds) that have been obtained from the conventional model reported in the previous section and a fully Bayesian model used for small area estimation. It appears that there is very good albeit not perfect agreement in terms of the parameter estimates. It should be noted however that perfect agreement would not be expected as these are not *exactly* the same model. The Bayesian model contains a random effect for each local authority.

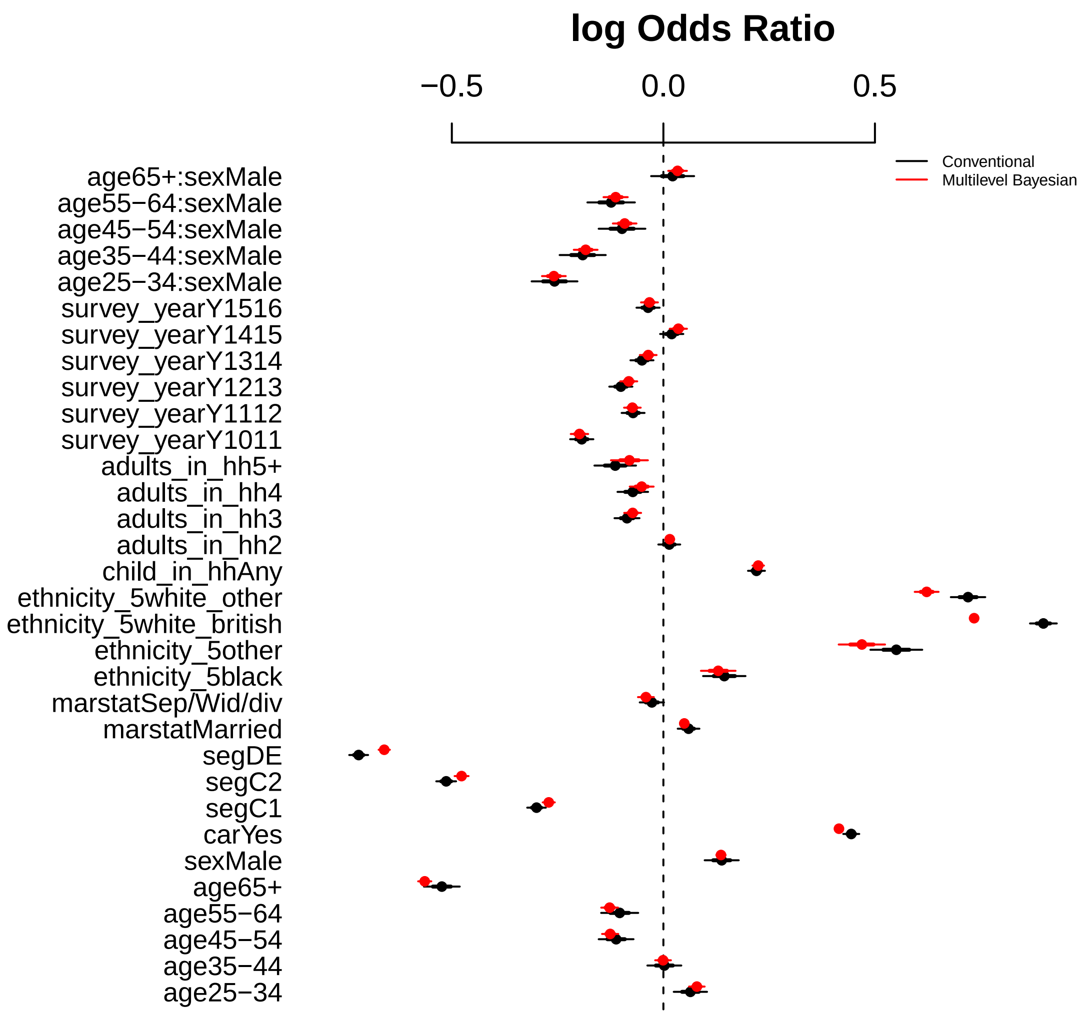

Figure 2.2: Question 1 (trips in last week): comparison of parameter ("fixed") estimates from frequentist (conventional) and multilevel Bayesian model fitted to Question 1

54. Figure 2.3 depicts a so-called “caterpillar plot” of the random effects for each of the local authorities obtained from the Bayesian model. This is a visual summary of the 95% posterior credible interval of the random effect in units of “log odds”. Each horizontal line represents a 95% credible interval for the “random effect” for each local authority. This represents the variation in outcomes that can be explained by a respondent coming from that authority. The names of the individual local authorities are not given in this plot for cosmetic reasons (they would be too small to read), however figure 2.4 will “zoom” in on a subset of authorities. The reason for providing this plot is to give some idea of the authority by authority random effects, and to indicate how much variation in trip taking could be explained by well chosen authority specific predictor variables. There are some data anomalies. In some cases, the local authority has not been identified, and the plot shows extremely wide intervals estimates. However, this picture is intended as a way of evaluating the typical width of a credible interval for any unmodelled local authority specific effect on the number of trips reported in the previous week.

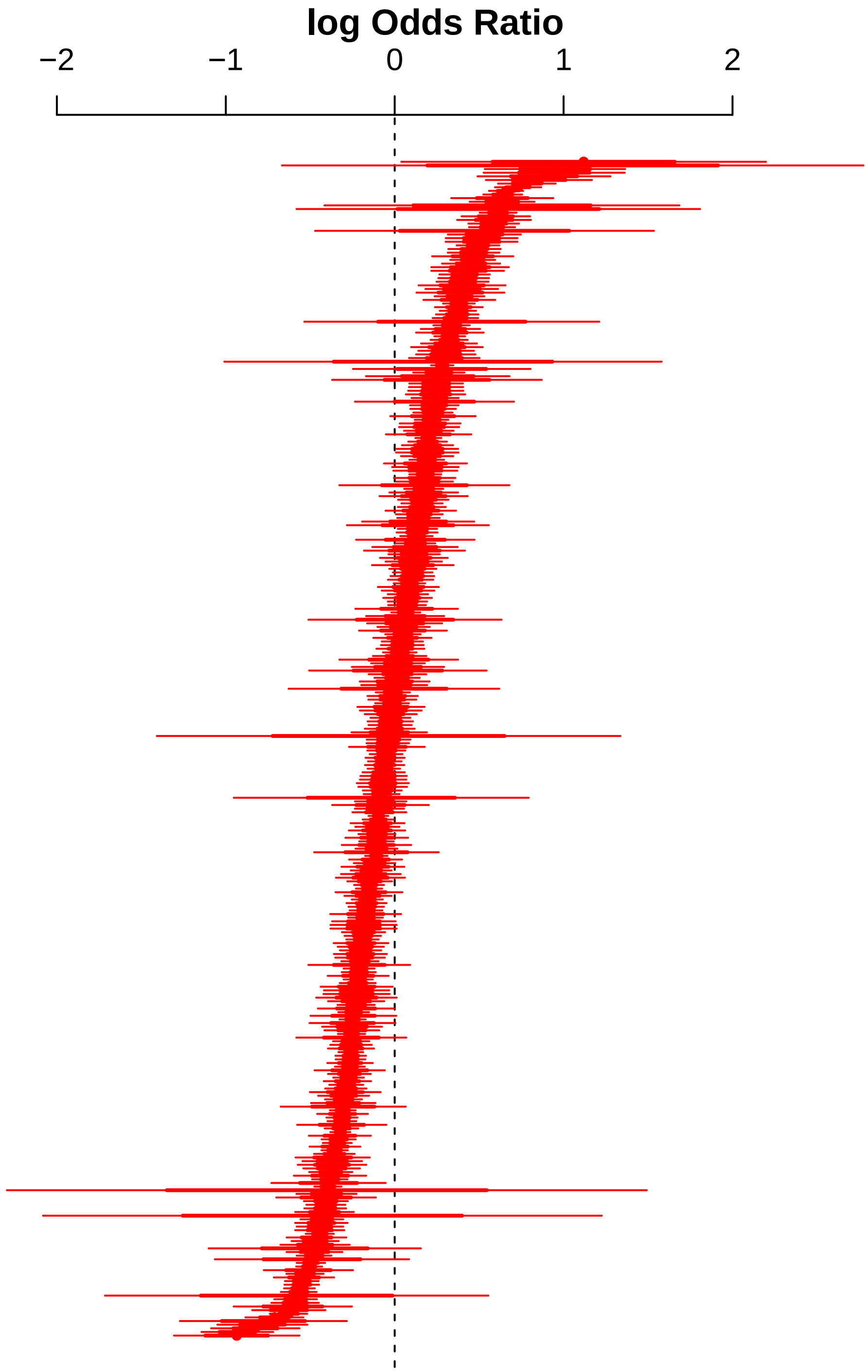

Figure 2.3: Question 1 (trips in last week): summary of interval estimates for local Authority Random effects from Bayesian model

55As a means of making the results more interpretable, figure 2.4 summarises the posterior credible interval for an *arbitrarily chosen* selection of local authorities[6]. The precision of the final small area estimates is a function of the width of these intervals. As noted in the previous section, it would be possible to model these in turn, using predictor variables that have social or geographic relevance and even to allow for spatial autocorrelation (to allow neighbouring areas to be similar in ways that are not explained by explicit numerical variables). The random effects for Exeter have interval estimates which are always above zero, whereas those for Hull have interval estimates which are always below zero. This tells us that a particular person type (of any combination of socio-demographic characteristics) is more likely than the average person to have taken a trip in the previous week if they were from Exeter and less likely if they were from Hull. These authority specific effects are well worth further exploration, both because they are of interest in their own right but also because developed a multilevel model at a relevant aggregate level should yield superior small area estimates by using relevant geographical information.

[6]

ONS code,	Name
E07000041	Exeter
E07000180	Vale of White Horse
E07000218	North Warwickshire
E07000122	Pendle
E06000010	Kingston upon Hull, City of
E09000027	Richmond upon Thames
E07000202	Ipswich
E07000205	Suffolk Coastal
E08000002	Bury
E07000028	Carlisle

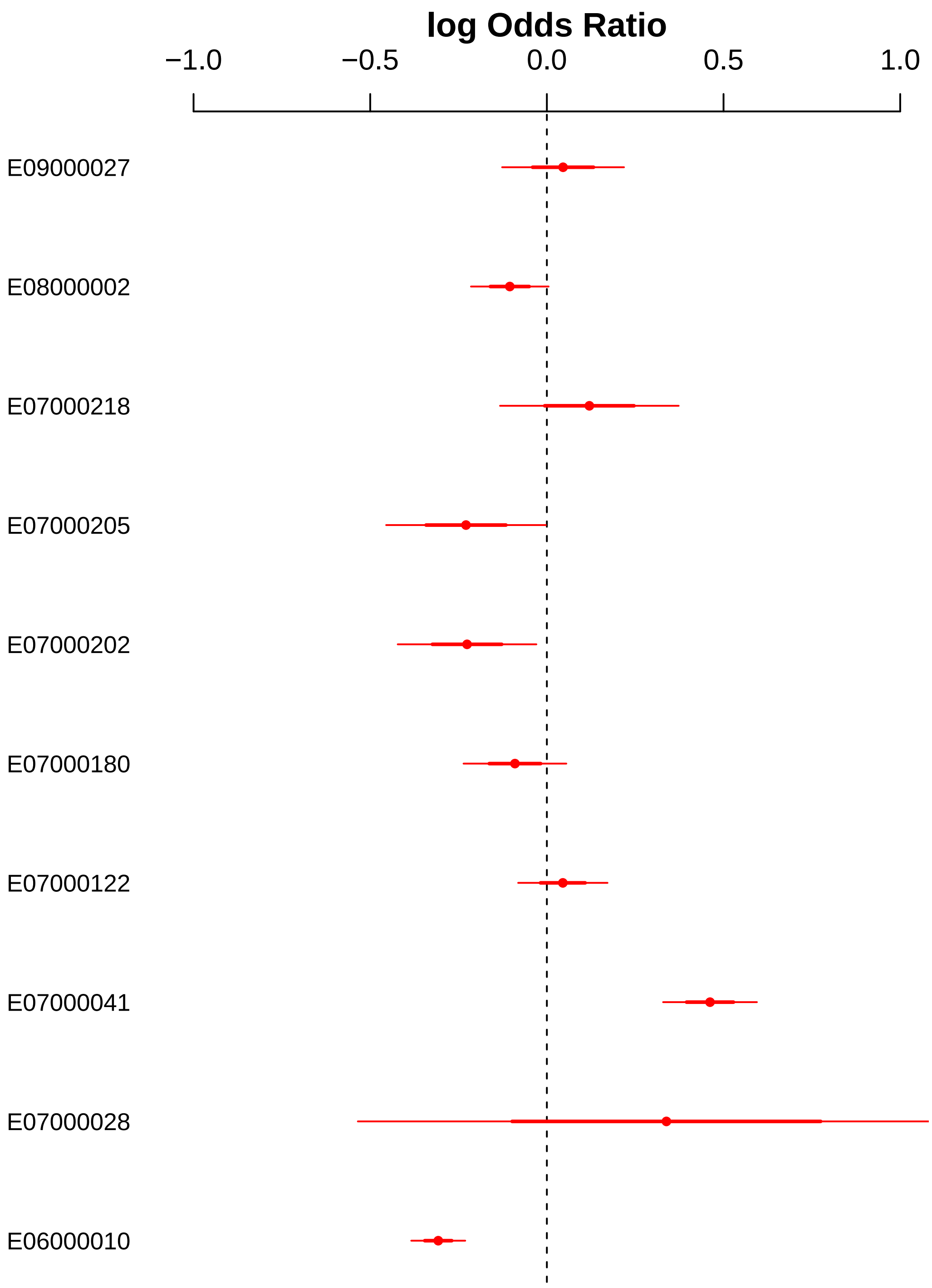

Figure 2.4: Question 1 (trips in previous week): exploded view of posterior random effects for some authorities

2.3.2 *Question 1 (trips in last week) 2015/2016 only*

56. In terms of evaluating the extent to which a reduced sample size, combined with small area methodology, might be feasible, results are next presented based solely on data from survey year 2015/16. Whilst better small area estimates can be obtained from using as many years data as possible, using data from a single year should give a more conservative (more pessimistic) guide to the effect of potentially reducing sample size for a single year.

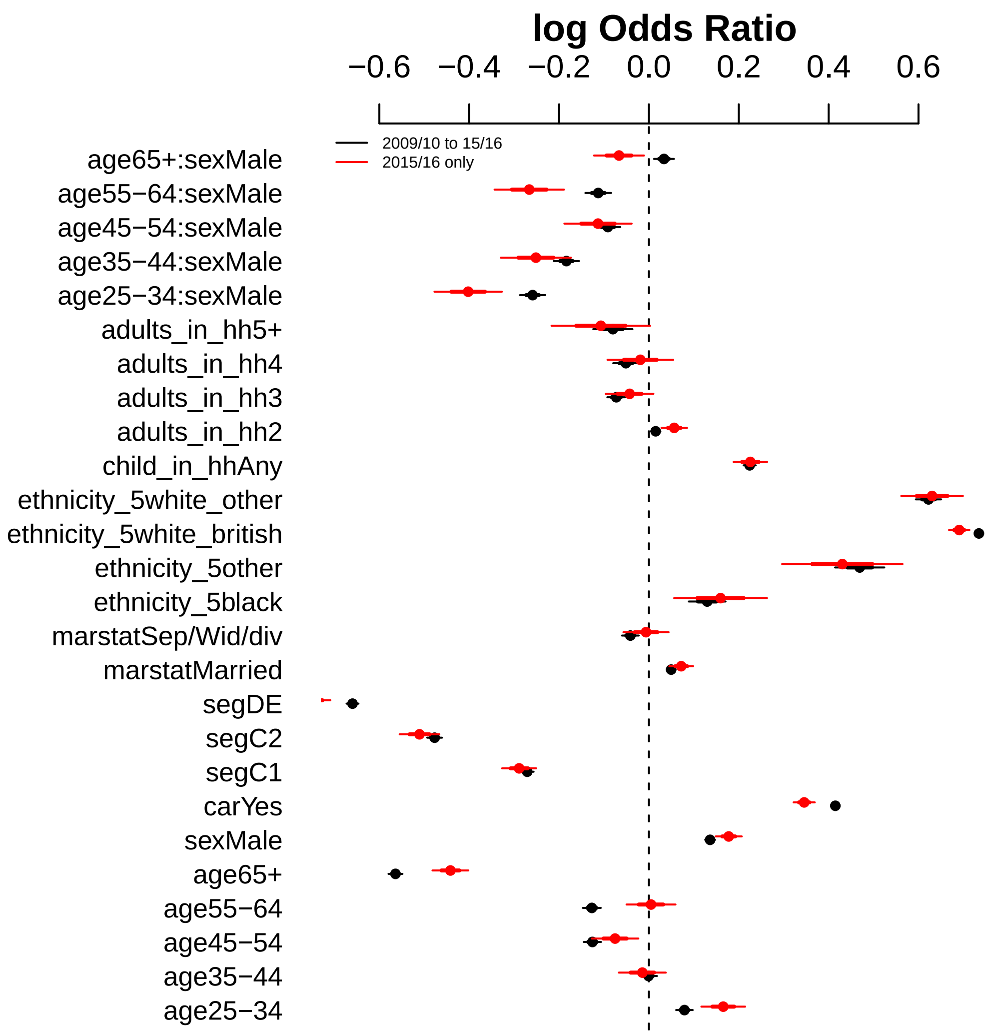

Figure 2.5: Question 1 (trips in last week): comparison of parameter ("fixed") estimates from modelling all years 2009/10 to 2015/16 and using data solely for 2015/16

57. The “fixed” effects are plotted in figure 2.5. It can be seen that, generally speaking, the estimates compare well.

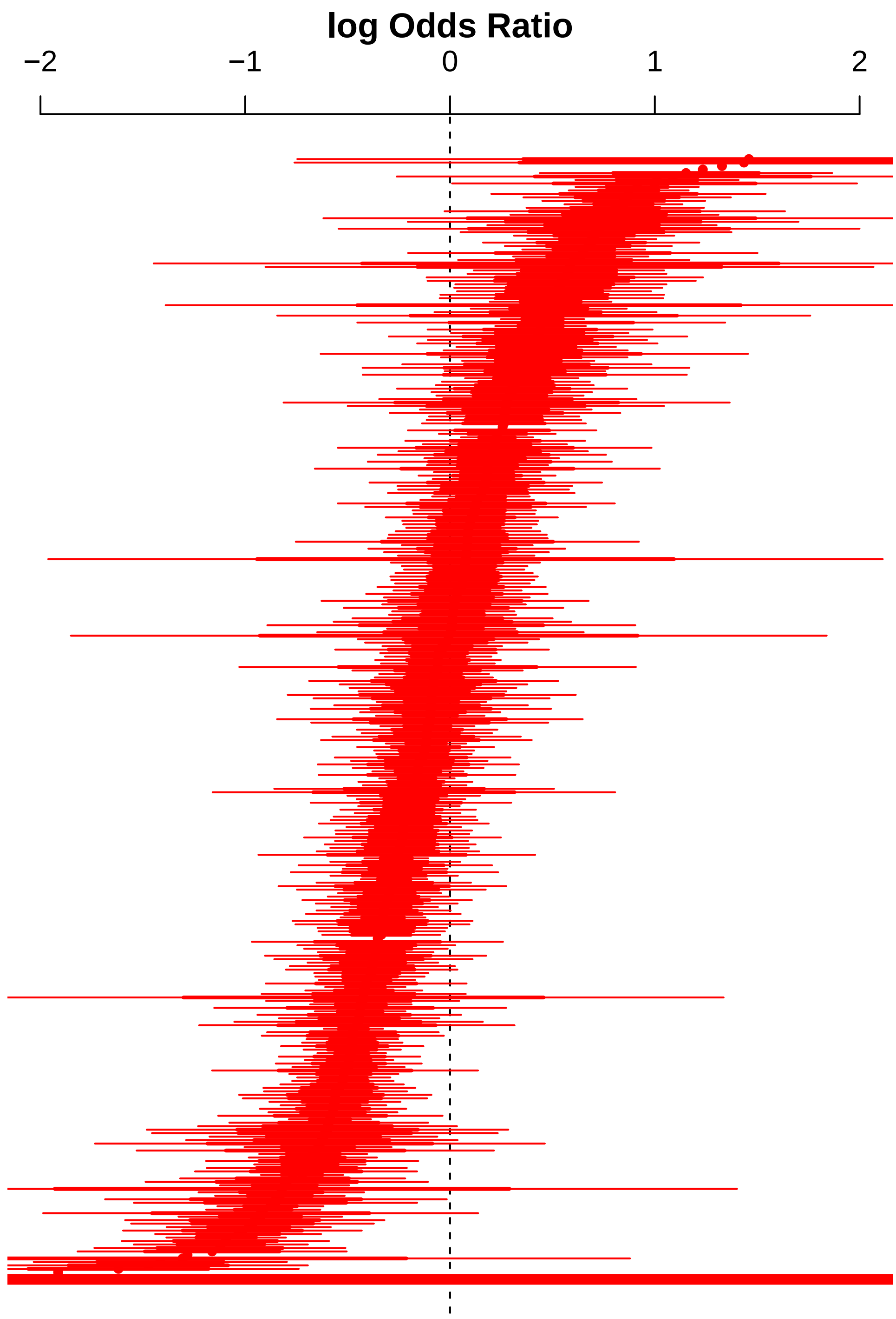

Figure 2.6: Question 1 (trips in previous week): local authority specific random effects for survey year March 2015 to February 2016

58 Again, a zoom is made available of the arbitrarily chosen authorities[7] Clearly as the subsamples reduce or remove respondents from particular local authorities, the width of the credible intervals can increase considerably.

[7]

ONS code,	Name
E07000041	Exeter
E07000180	Vale of White Horse
E07000218	North Warwickshire
E07000122	Pendle
E06000010	Kingston upon Hull, City of
E09000027	Richmond upon Thames
E07000202	Ipswich
E07000205	Suffolk Coastal
E08000002	Bury
E07000028	Carlisle

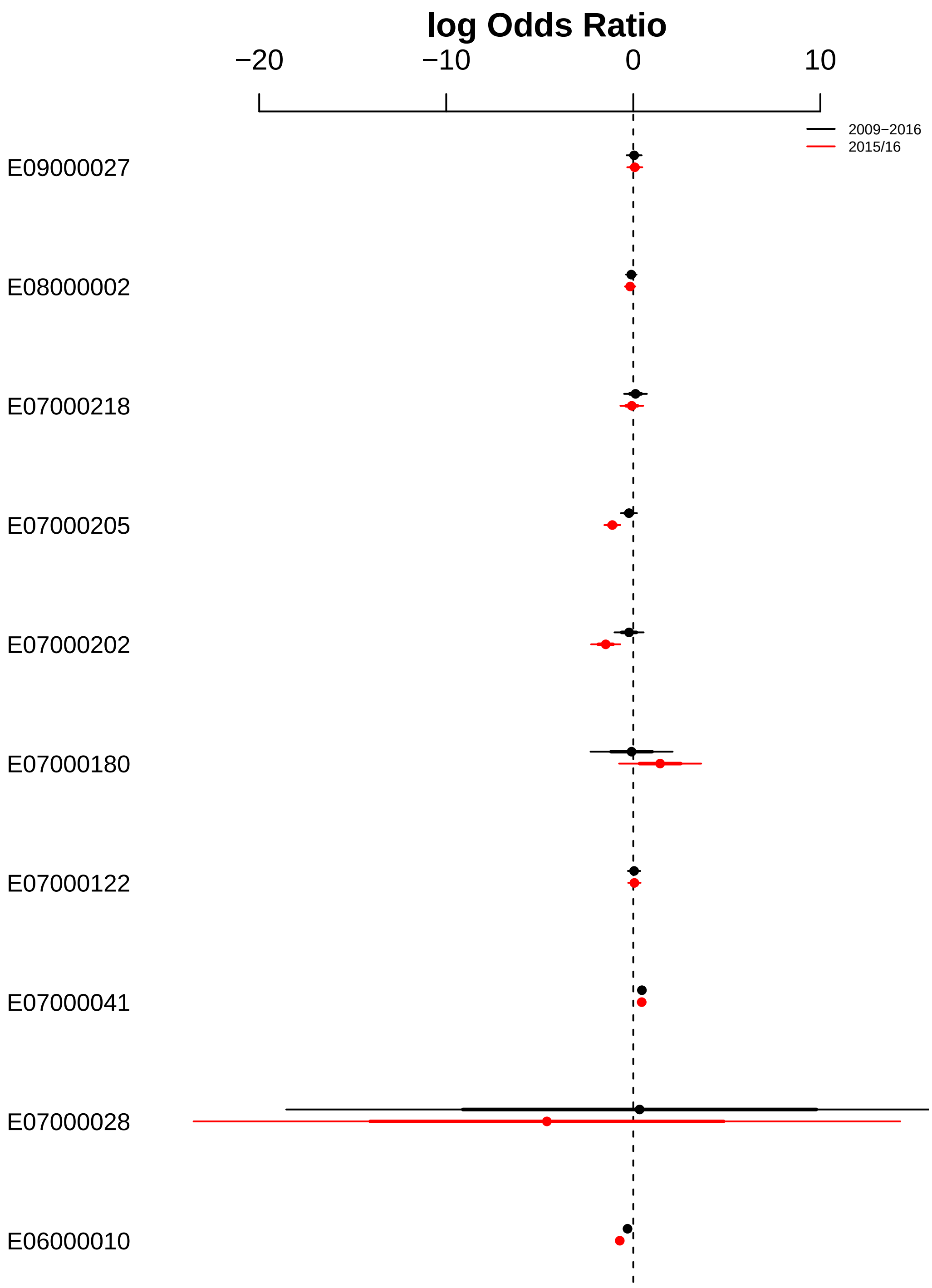

Figure 2.7: Question 1 (trips in previous week): expanded view of some local authority random effects from 2017/2018 survey year

2.3.3 *Exploring the effect of reduced sample size*

59 A key outcome of this feasibility study was to determine whether small area methods could allow a reduced sample size in future use. To explain the impacts of reduced sample size, an illustrative study on the modelling is presented. A random sub-sample of the 2015/2016 data are taken. Results presented here show an 80% sample and then a 50% sample. It can be seen that the intervals are widening with the 50% sample.

60. Figure 2.8 compares the “fixed” effects estimates from the models fitted to all the 2015/16 data, an 80% sample and a 50% sample. As the interval width increases, the precision of any small area estimates decreases.

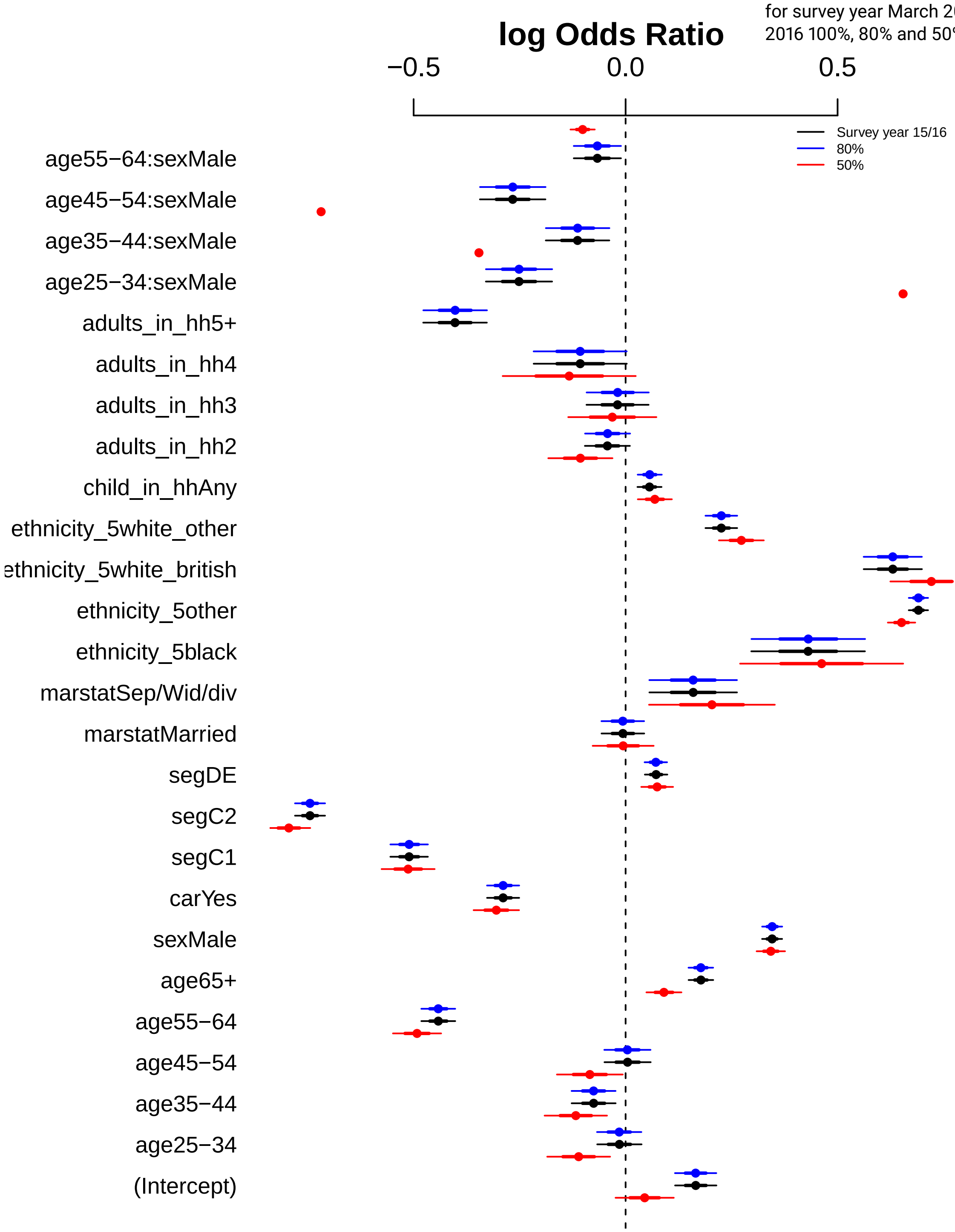

Figure 2.8: Question 1 (trips in last week): comparison of "fixed" effects for survey year March 2015 to February 2016 100%, 80% and 50% subsample

61. Figures 2.20, 2.21 and 2.22 show the “caterpillar” plots for the random effects based on all data, an 80% sub-sample and a 50% sub-sample respectively. Again, a “zoomed” caterpillar plot of the posterior random effects for the ten arbitrarily chosen local authorities is presented in figure 2.12

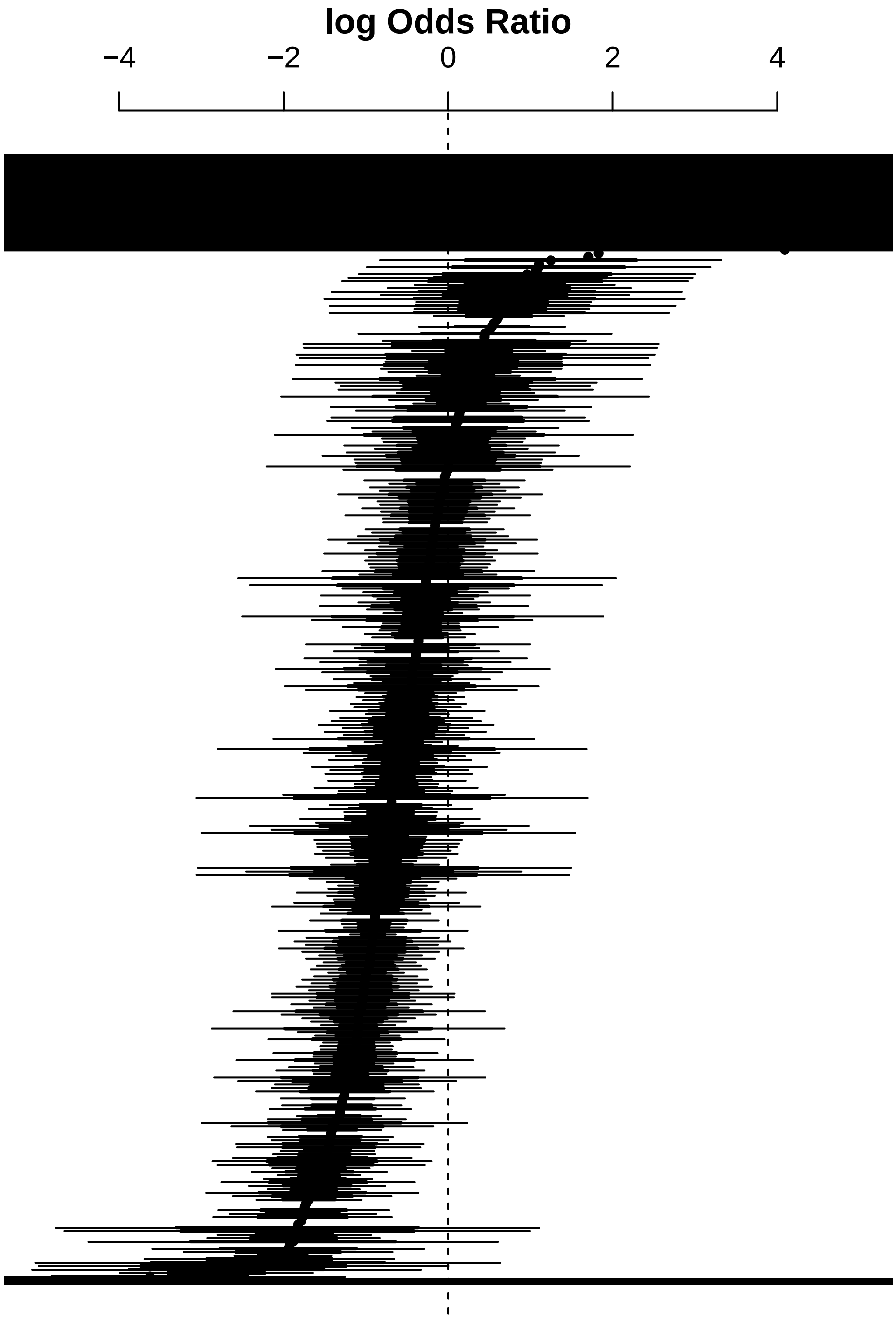

Figure 2.9: Q1 (trips in last week): local authority specific random effects for survey year March 2017 to February 2018 based on all data

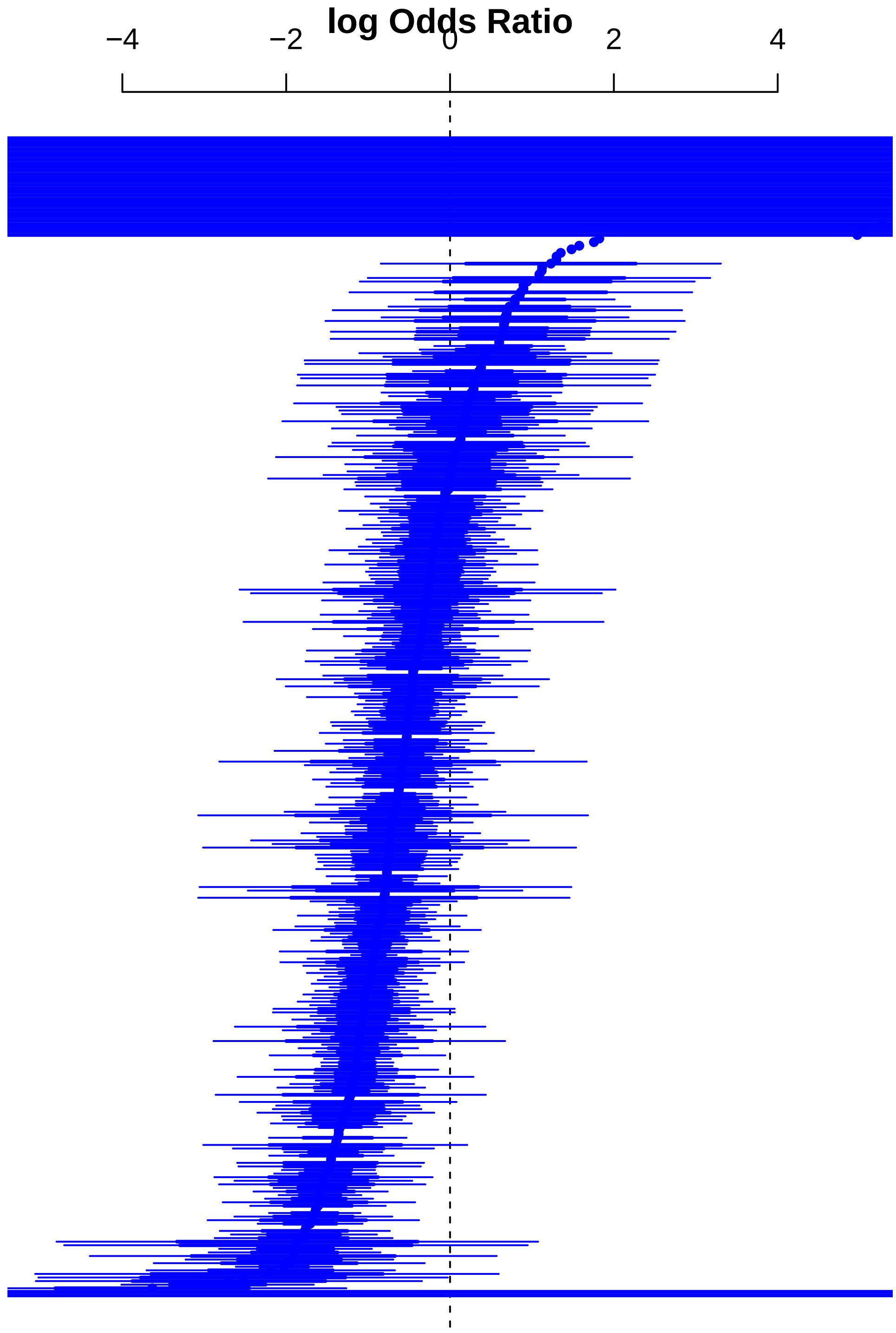

Figure 2.10: Question 1 (trips in last week): local authority specific random effects for survey year March 2015 to February 2016 based on 80% subsample

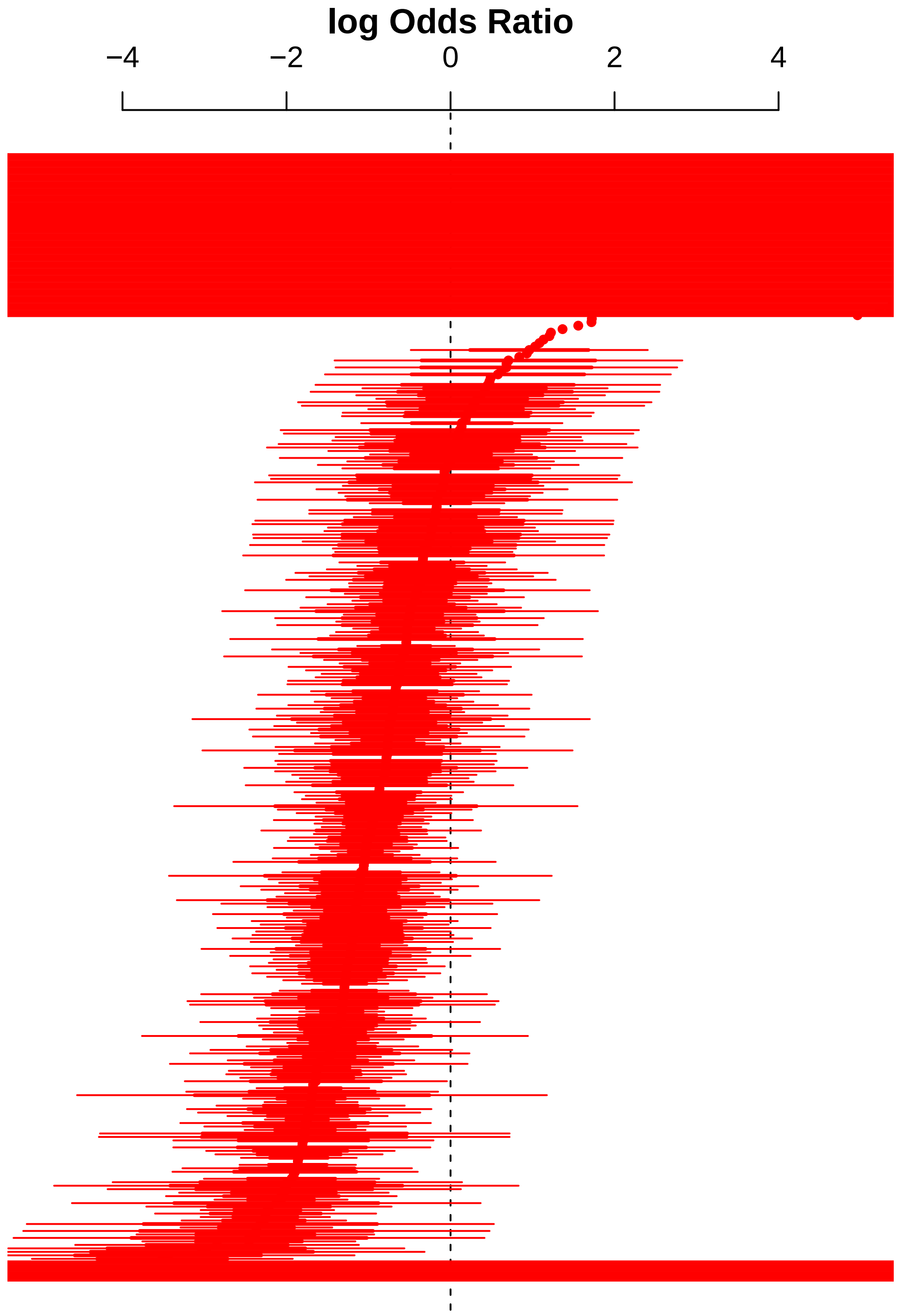

Figure 2.11: Question 1 (trips in previous week): local authority specific random effects for survey year March 2015 to February 2016 based on 50% subsample

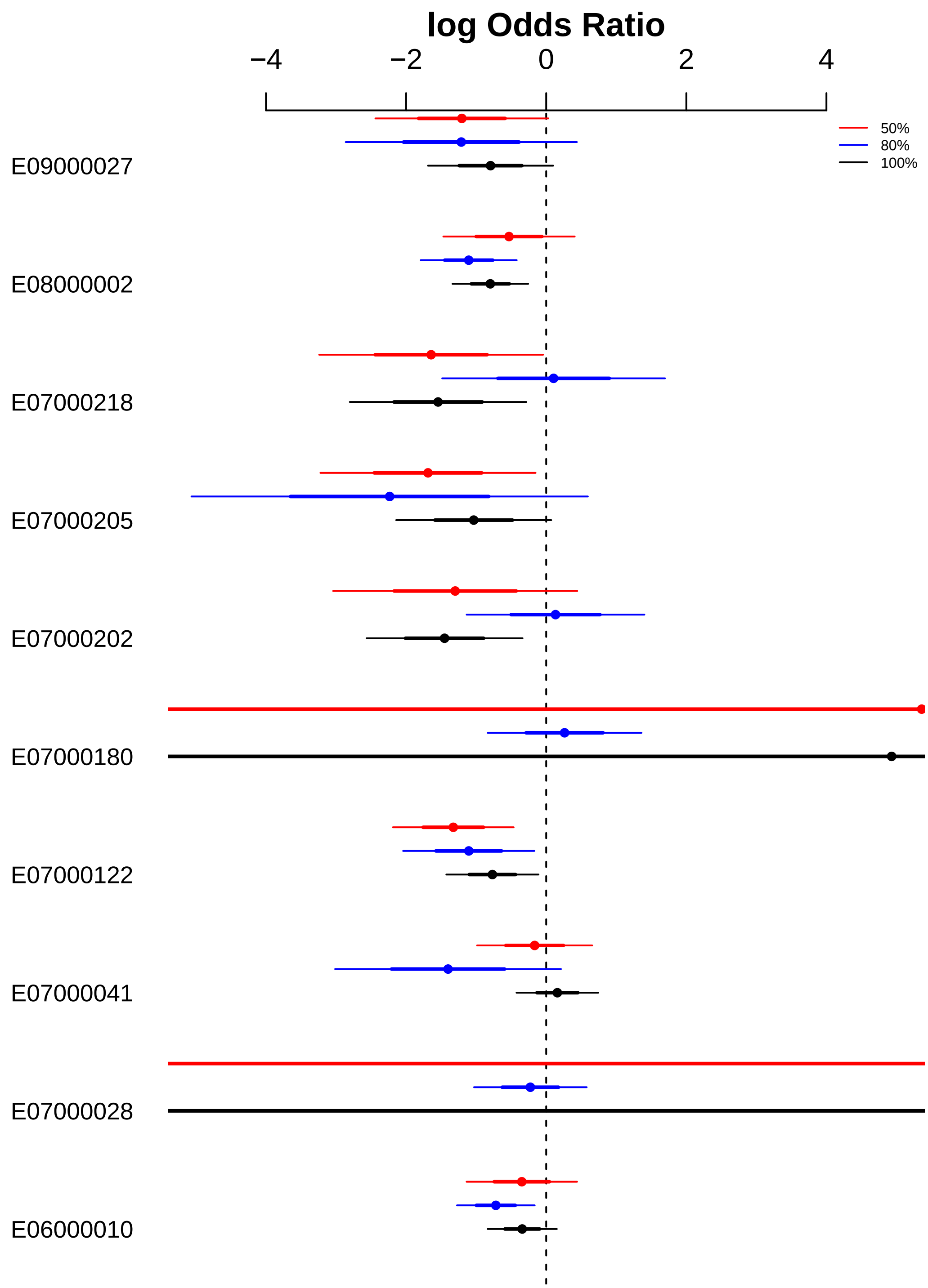

Figure 2.12: Question 1 (trips in last week): expanded view of selected local authority random effects from 2017/2018 survey year under all, 80% and 50% sampling

62Finally, it is illustrative to note the sample size attained by these scenarios. This is given in table 2.1.

Sample	Number of rows of data
2009/10 - 2015/16	326755
2015/16 100% sample	45965
2015/16 80% sample	36772
2015/16 50% sample	22982

Table 2.1: Sample size (number of data point from MENE survey used to fit models to Q1

2.3.4 Q17 (at least monthly trips in the last year): Bayesian modelling and reduced sample size

63. Similar results are briefly presented for Question 17 to those given earlier for Question 1.

64. As before, the first set of plots summarise the results of fitting a plot to the entire data set from survey year 2009/2010 through to 2015/2016. Figure 2.16 contrasts the conventional model reported in the previous section and a fully Bayesian model used for small area estimation. As before, it appears that there is good but not perfect agreement in terms of the parameter estimates.

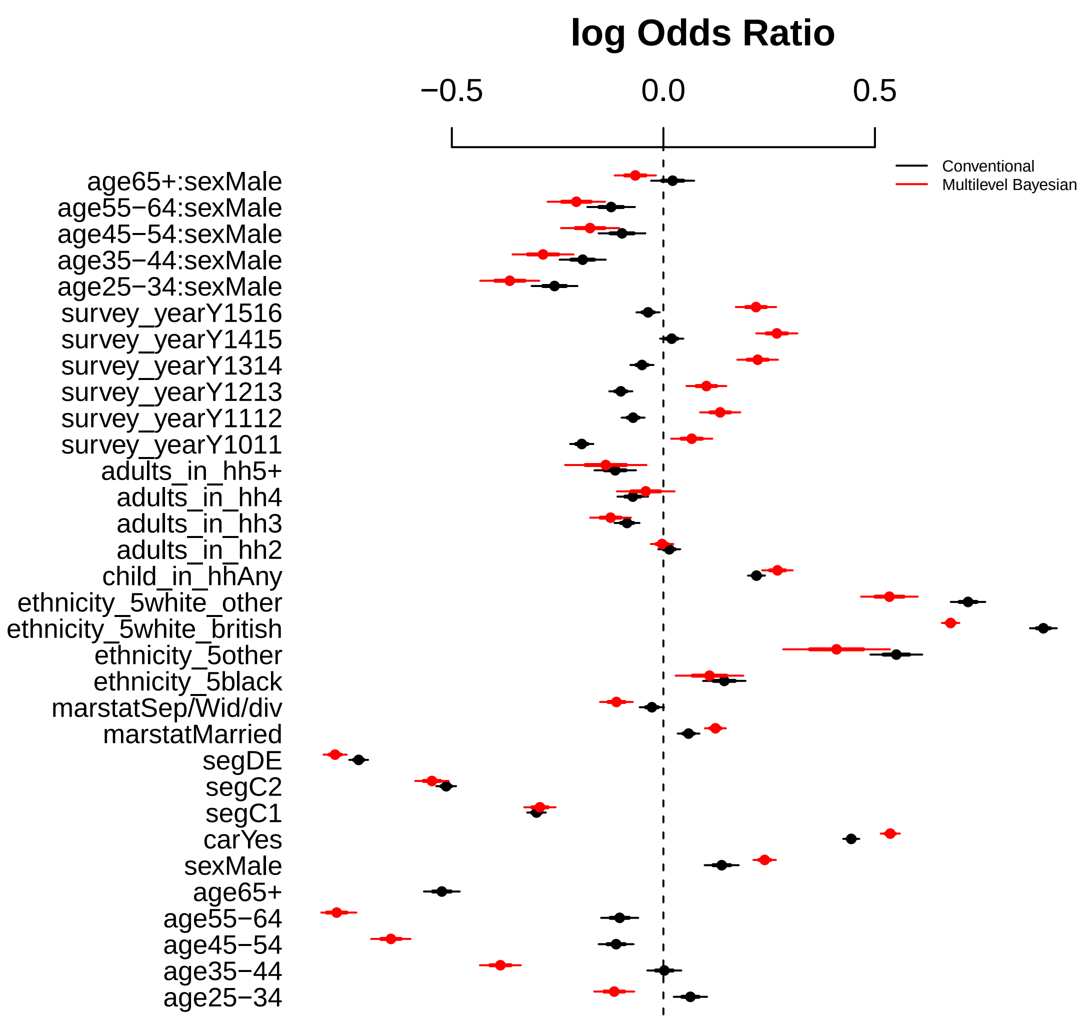

Figure 2.13: Question 17 (previous year, more than monthly trips): comparison of parameter ("fixed") estimates from frequentist (conventional) and multilevel Bayesian model fitted to Question 1

65. Figure 2.14 illustrates the “caterpillar plot” of the random effects for each of the local authorities obtained from the Bayesian model. and figure 2.15 provides a “zoom” view of the subset of authorities.

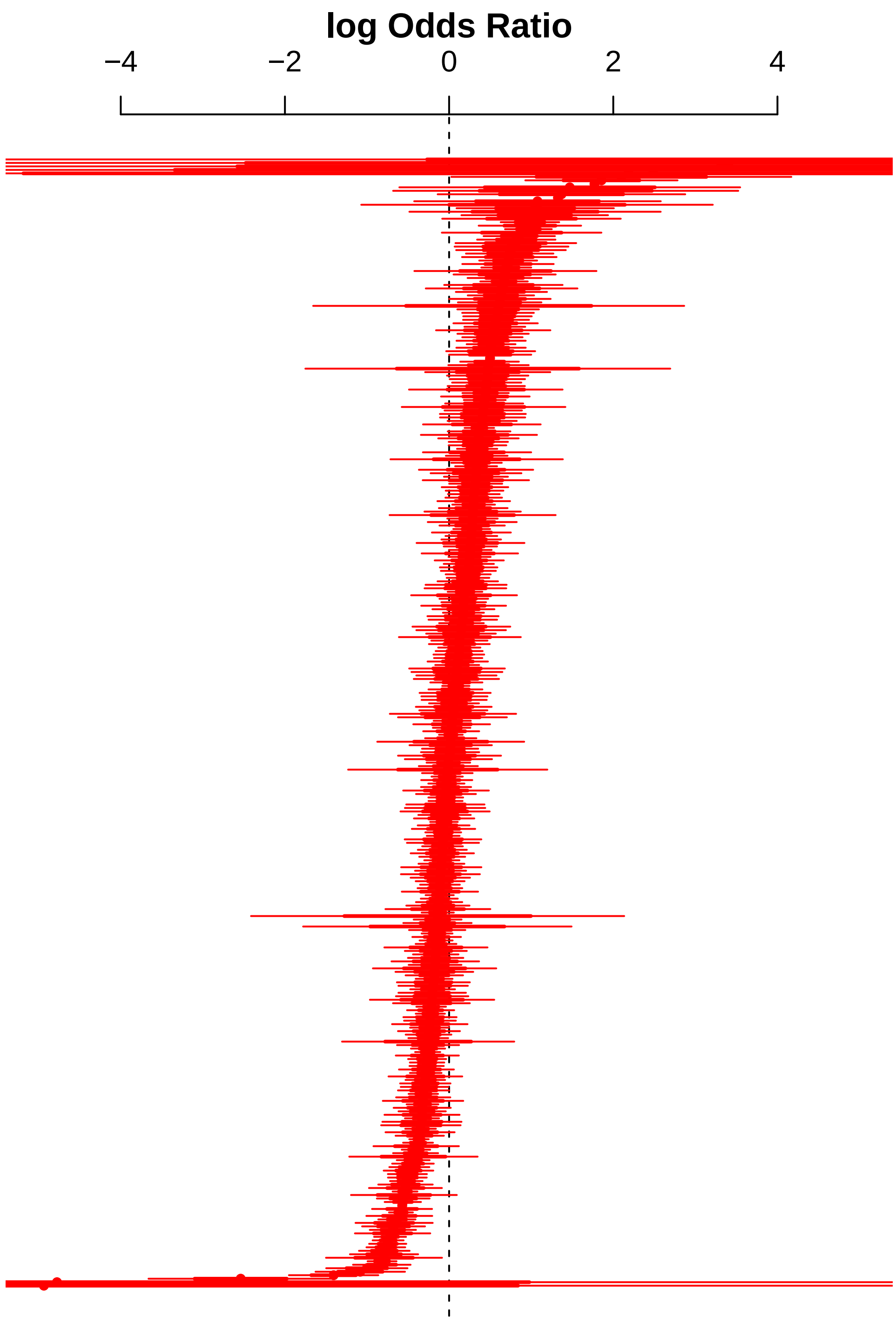

Figure 2.14: Question 17 (previous year, more than monthly trips): summary of interval estimates for local Authority Random effects from Bayesian model

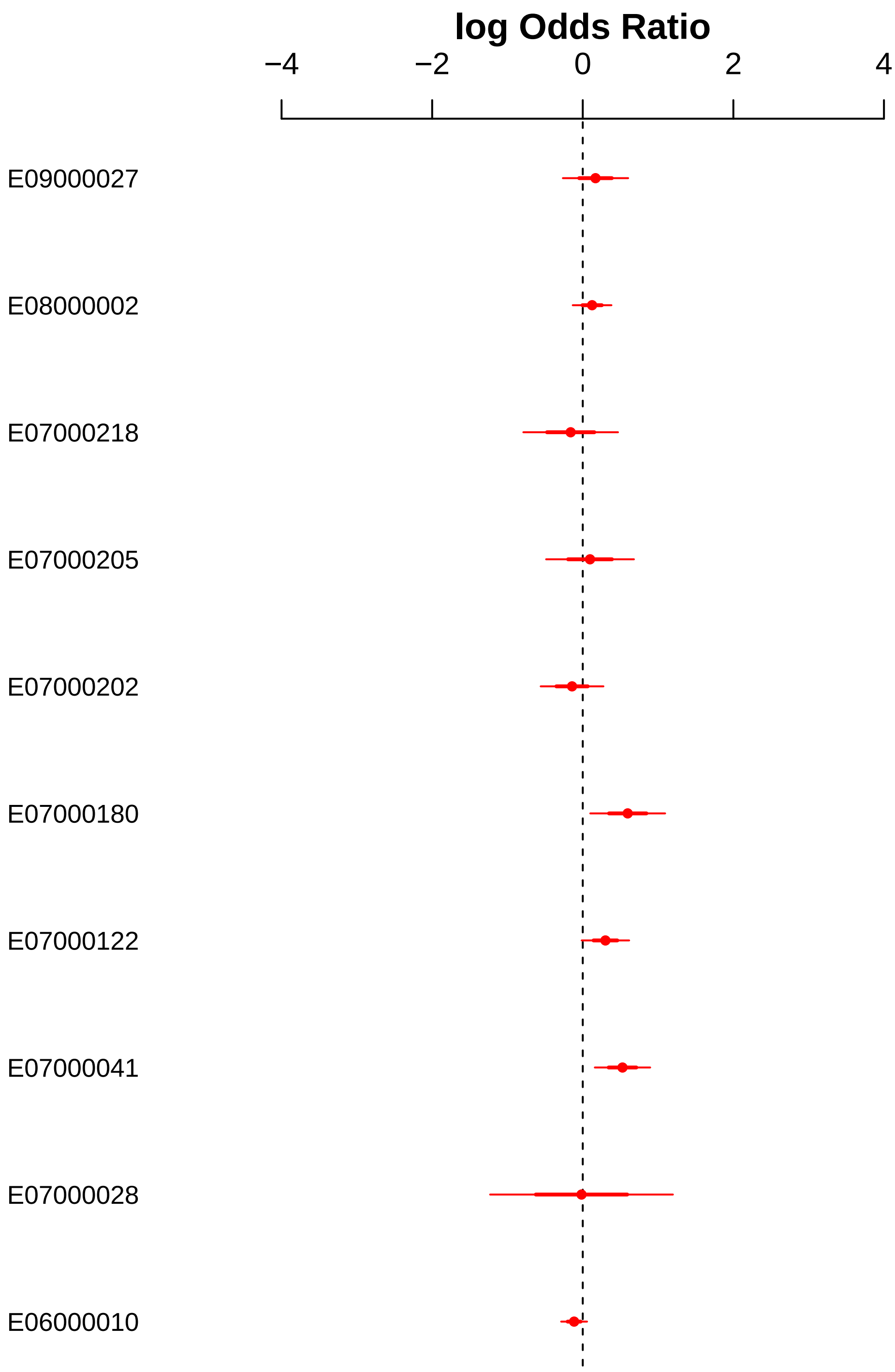

Figure 2.15: Question 17 (previous year, more than monthly trips): exploded view of posterior random effects for some authorities

2.3.5 *Question 17 (previous year, more than monthly trips) 2015/2016 only*

66For small area predictions, as before, results are based solely on data from survey year 2015/16.

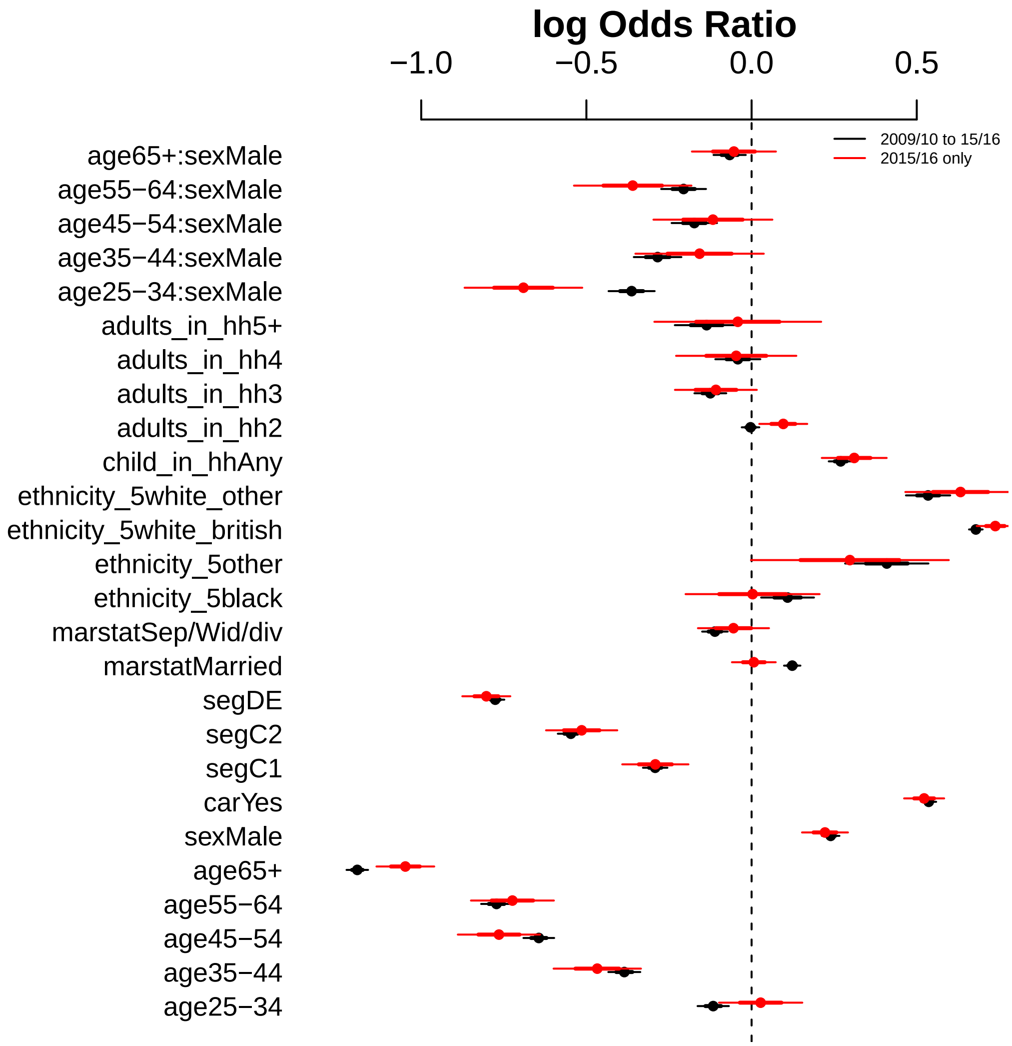

Figure 2.16: Question 17 (previous year, more than monthly trips): comparison of parameter ("fixed") estimates from modelling all years 2009/10 to 2015/16 and using data solely for 2015/16

67. Figure 2.17 depicts the random effects, and figure 2.18 shows the exploded view for the same arbitrarily chosen authorities as before for a model fitted to the 2015/16 only data.

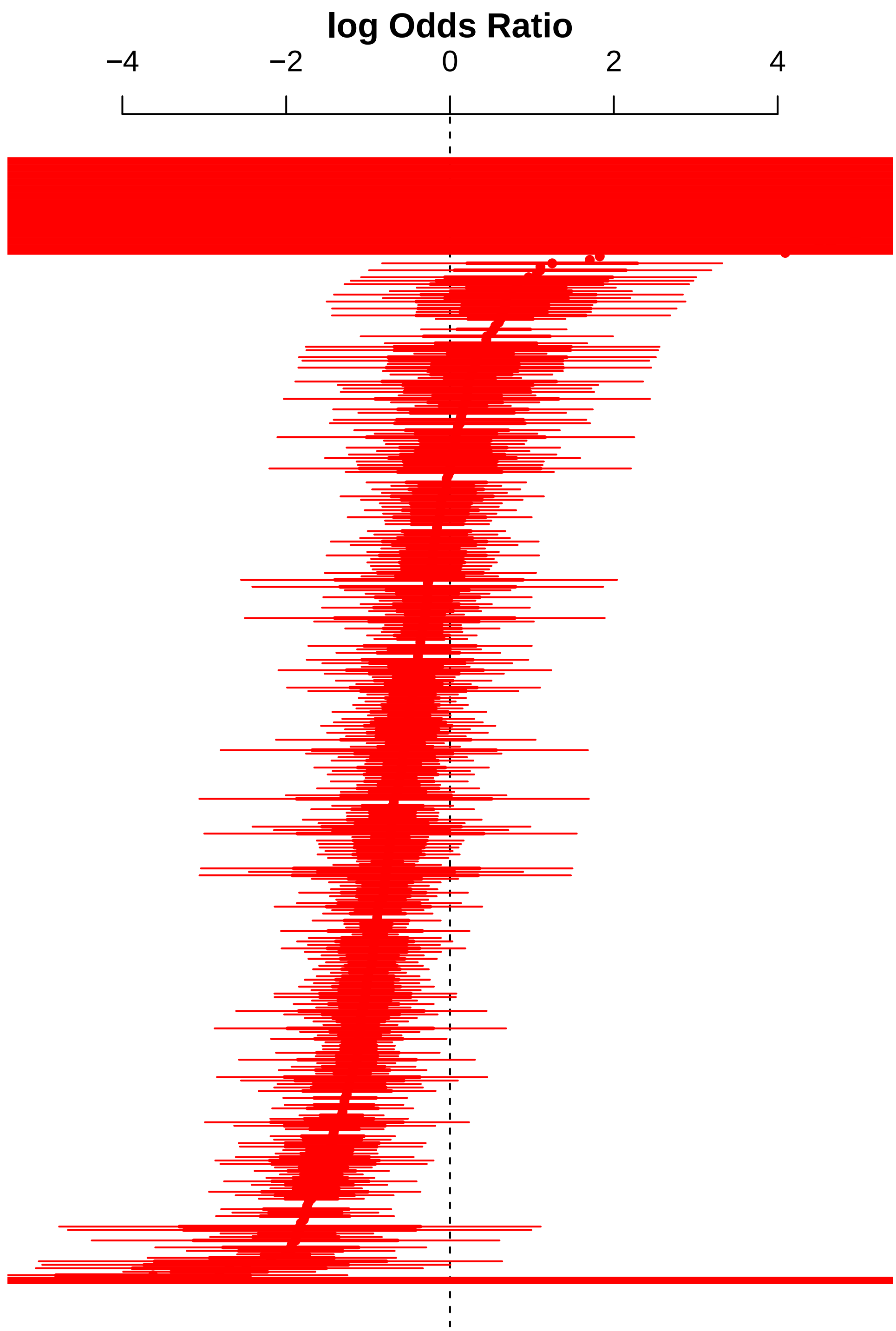

Figure 2.17: Question 17 (previous year, more than monthly trips): local authority specific random effects for survey year March 2017 to February 2018

2.3.6 *Exploring the effect of reduced sample size*

68. For illustration purposes, a random sub-sample of the 2015/2016 data are taken. Results presented here show an 80% sample and then a 50% sample. Figure 2.19 compares the “fixed” effects estimates from the models fitted to all the 2015/16 data, an 80% sample and a 50% sample. As the interval width increases, the precision of any small area estimates decreases. It can be seen that the intervals are widening with the 50% sample.

69. Figure 2.8 compares the “fixed” effects estimates from the models fitted to all the 2015/16 data, an 80% sample and a 50% sample. As the interval width increases, the precision of any small area estimates decreases.

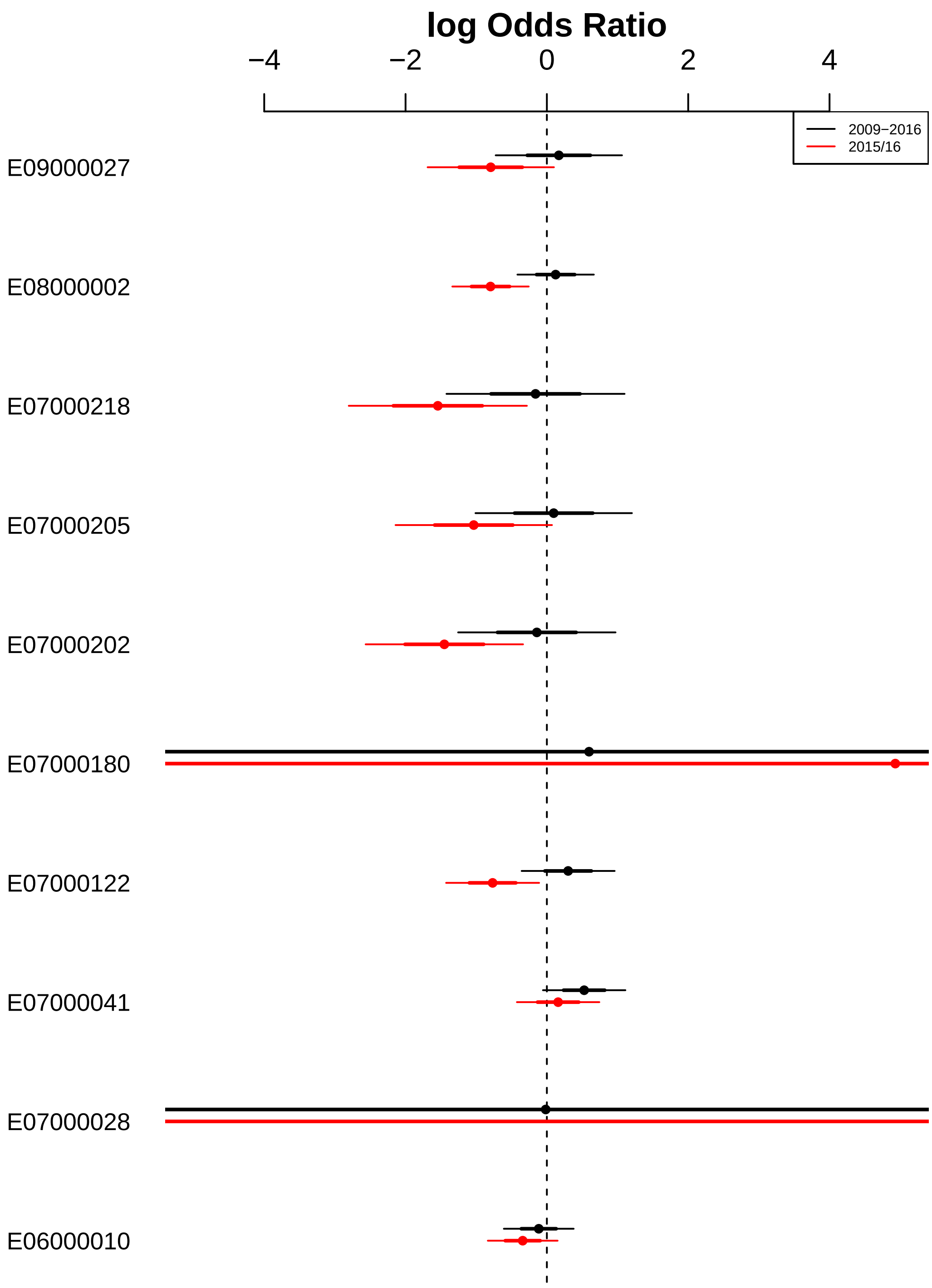

Figure 2.18: Question 17 (previous year, more than monthly trips): expanded view of some local authority random effects from 2015/2016 survey year

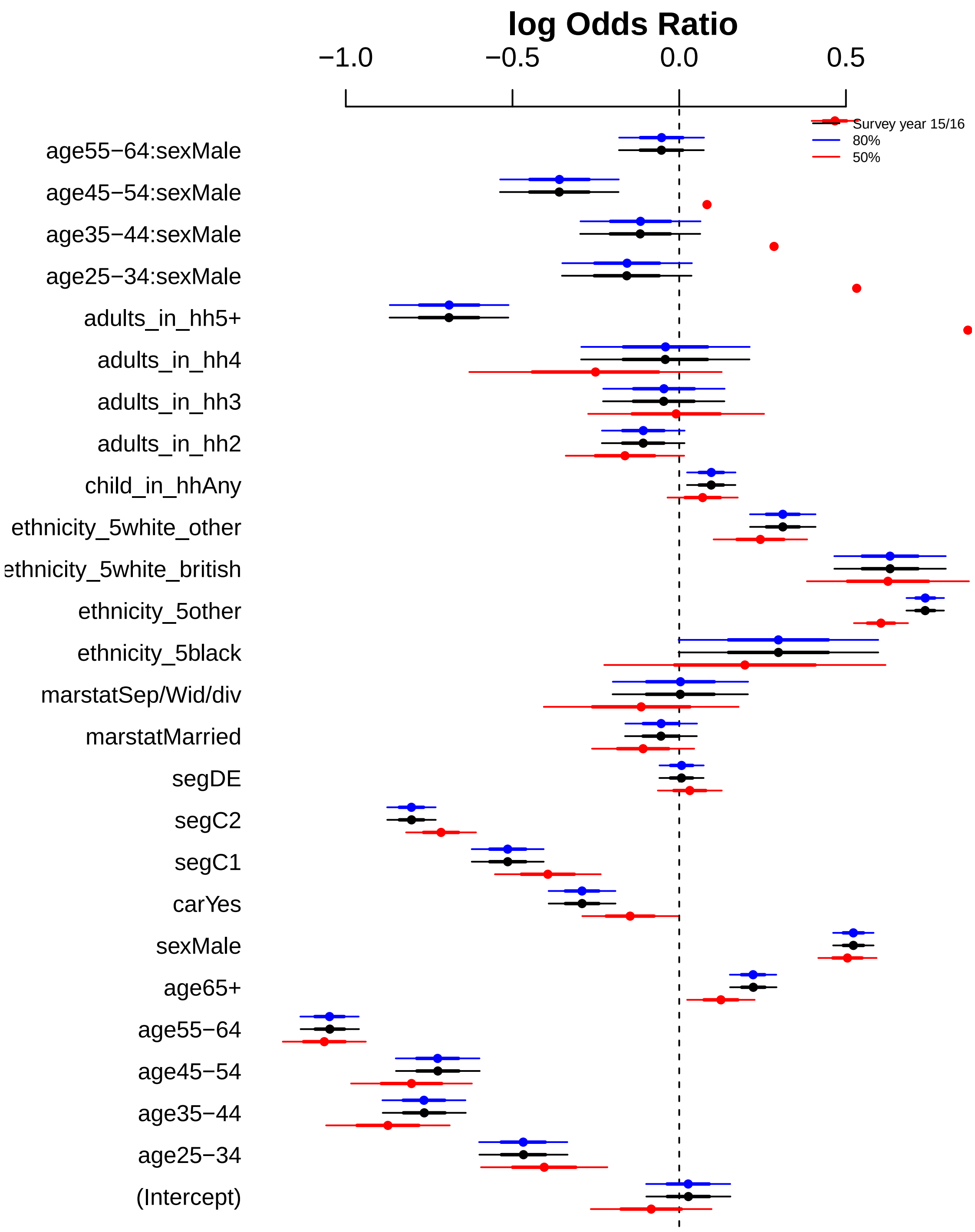

Figure 2.19: Question 17 (previous year, more than monthly trips): fixed effects estimates from 100%, 80% and 50% subsamples from 2015/16

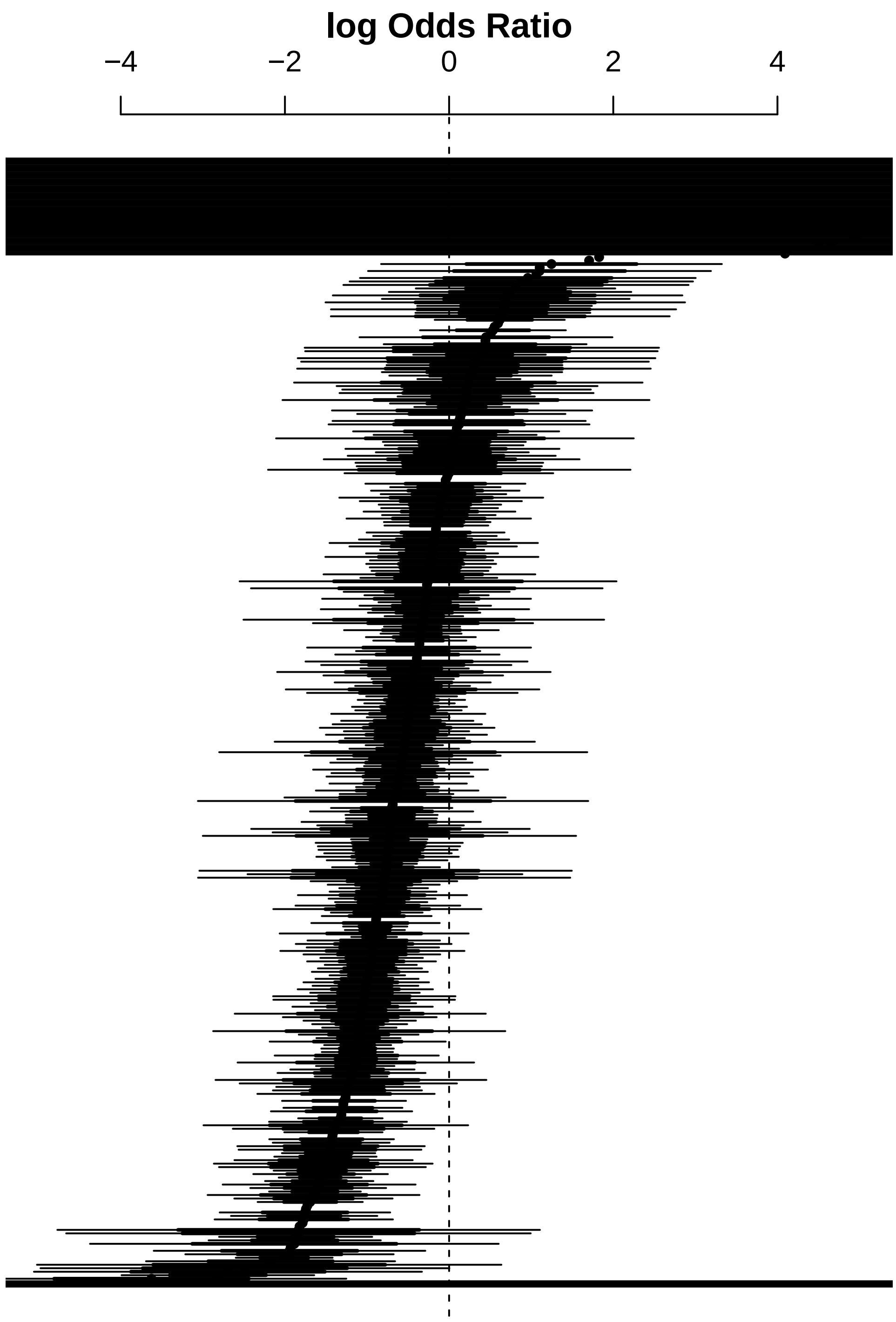

Figure 2.20: Q17 (previous year, more than monthly trips): local authority specific random effects for survey year March 2017 to February 2018 based on all data

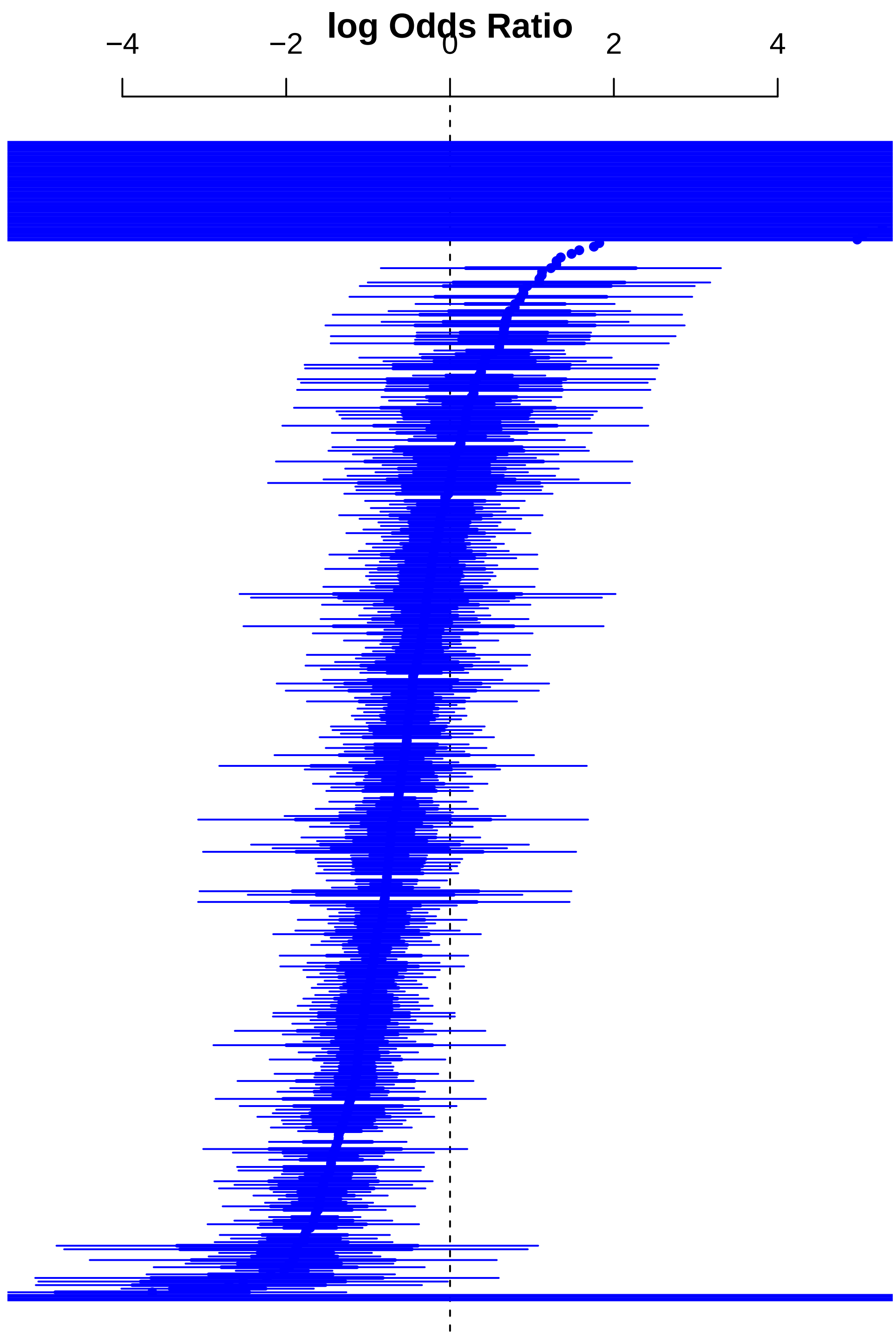

Figure 2.21: Question 17 (previous year, more than monthly trips): local authority specific random effects for survey year March 2015 to February 2016 based on 80% subsample

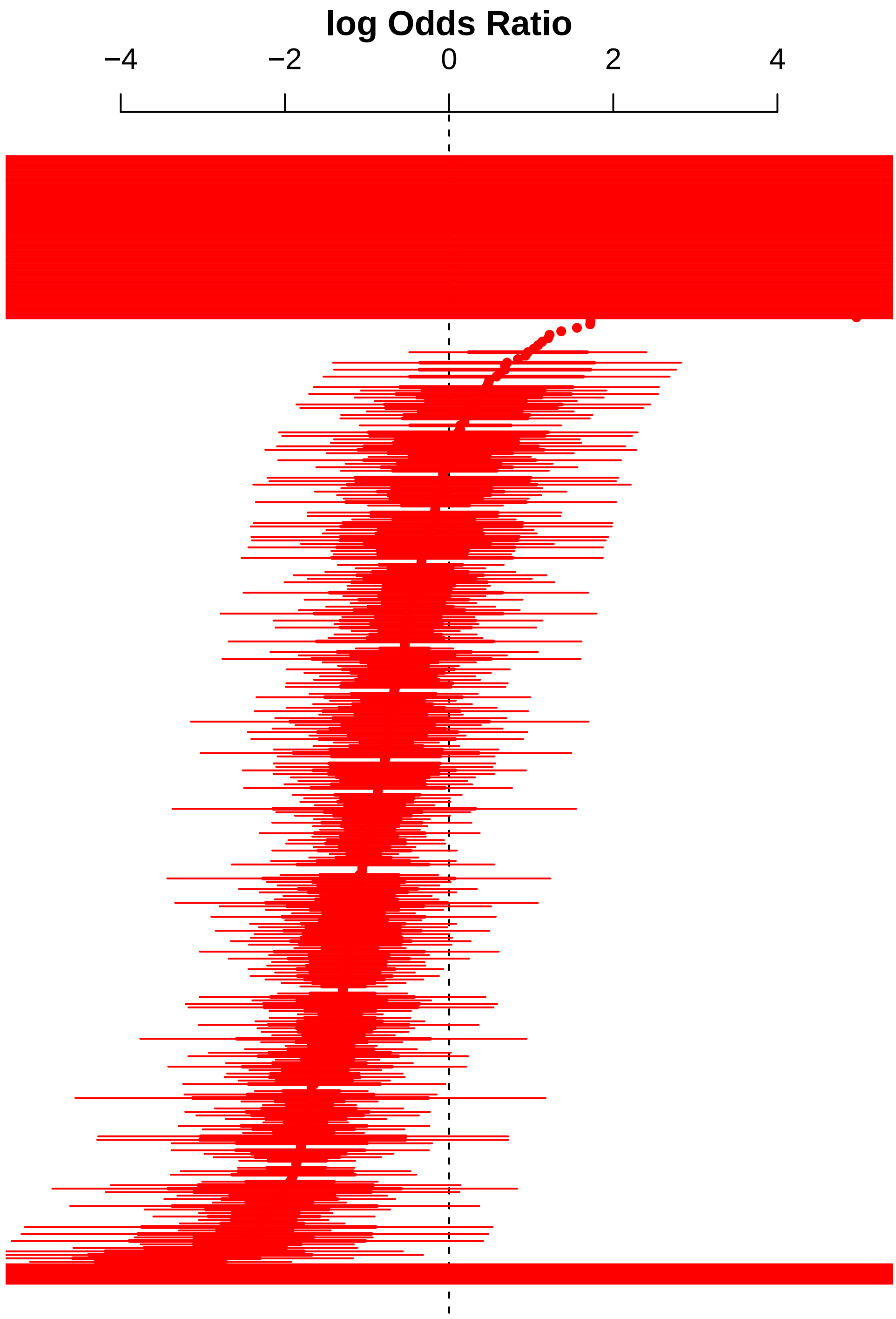

Figure 2.22: Question 17 (previous year, more than monthly trips): local authority specific random effects for survey year March 2015 to February 2016 based on 50% subsample

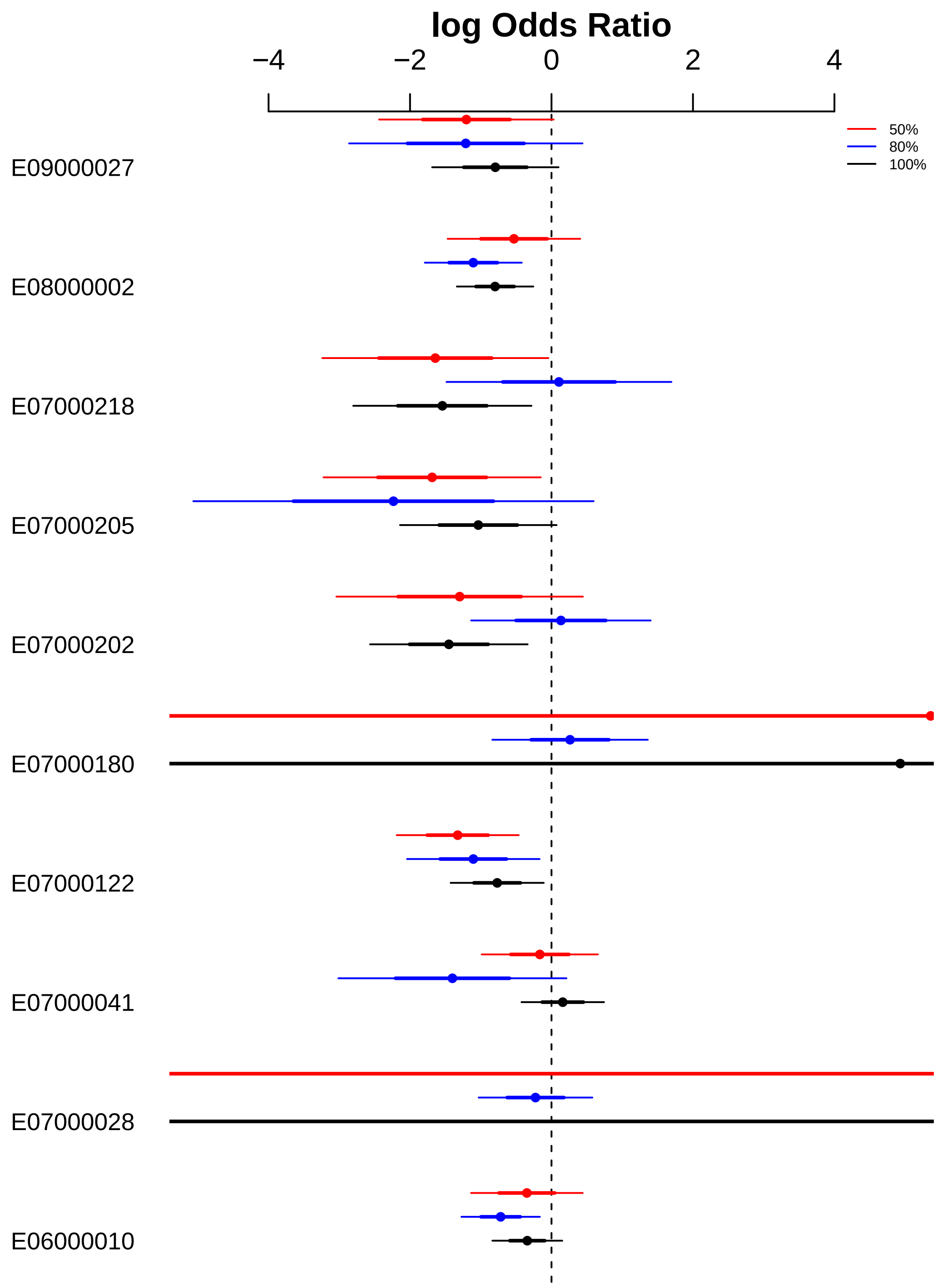

Figure 2.23: Question 17 (previous year, more than monthly trips): expanded view of selected local authority random effects from 2017/2018 survey year under all, 80% and 50% sampling

70. Finally, the sample sizes on the which the Question 17 model fits are based are given in table 2.2

Sample	Number of rows of data
2009/10 - 2015/16	74567
2015/16 100% sample	10676
2015/16 80% sample	8558
2015/16 50% sample	5291

Table 2.2: Sample size (number of data point from MENE survey used to fit models to Q17 (previous year, more than monthly trips)

71. An extensive study has been conducted whereby small area predictions are created in the form of a predicted count for each local authority. The process is as follows:

1. Small area predictions are generated for each individual in the SAM data as described in figure 2.1 above.
2. The SAM data are a 5% sample of the population. The small area predictions for each authority are collapsed to a count in each age-sex band.
3. The ratio of the number of respondents in the 2019 ONS population projection for each authority-age-band to the number of respondents in each SAM authority-age-band is used as a weight.
4. The small area predictions from the SAM are then upweighted using this ratio to generate a count for each local authority.

This process is repeated for a range of scenarios, where a random subsample is taken of the 2015/16 MENE survey data is taken. The Residual Mean Square Error, calculated as:

$$RMS = \sqrt{\frac{\sum_{i-1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{2.3}$$

where y_i denotes the small area estimated number of individuals taking a trip in a given local authority, and $\hat{y}$ denotes the estimated number of individuals based on a random subsample.

Sample size relative to full dataset	Residual Mean Square Error
95% Sample	1528.28
80% Sample	2918.92
75% Sample	3031.13
50% Sample	3788.14
30% Sample	7149.77
10% Sample	11672.91

Table 2.3: Q1: Root mean square error of sub-sample predictions

72. Clearly it is not ideal to use a small area estimate itself to calibrate the RMS, but this may be suitable for internal use and suggests that an 80% sample can be taken with little substantial increase in RMS error.

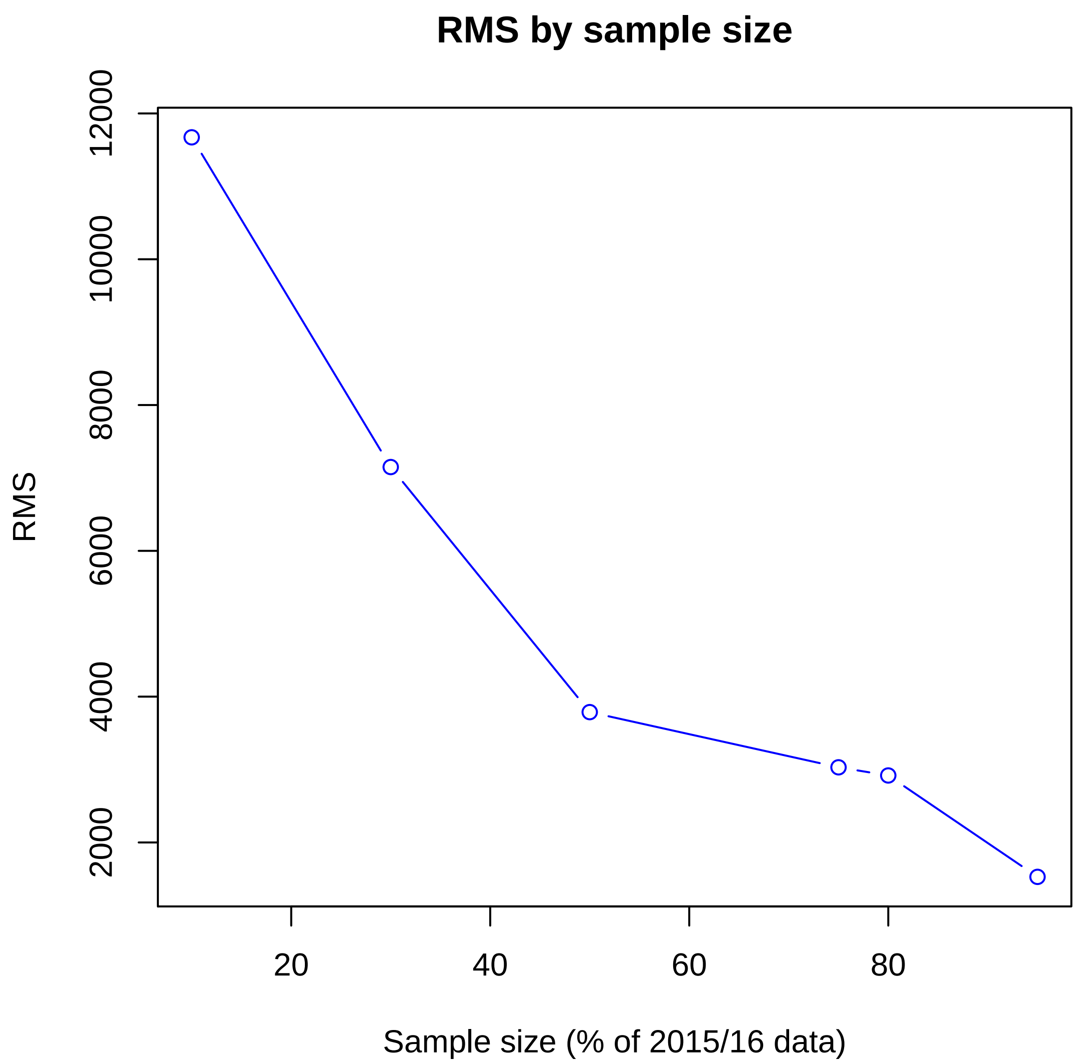

Figure 2.24: Q1 (trips in last week): estimates Residual Mean Square for random sub-samples relative to whole dataset for 2015/16

73. Whilst full results are given in a `csv` file, figure 2.25 presents a choropleth map of the number of people in each authority area who had made a trip to the natural environment in the previous week.

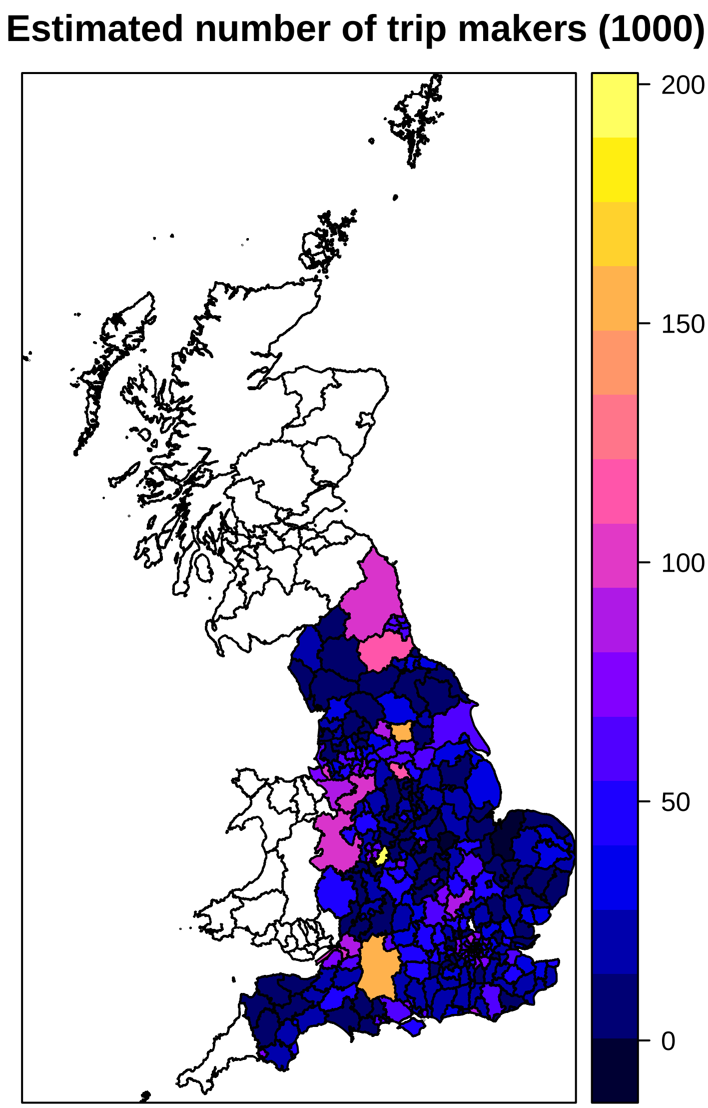

Figure 2.25: Small area estimate of the number of people in each lower level authority reporting a trip to the natural environment based on 2015/16 MENE data

2.3.7 *Question 17: effect of reducing sample size*

74. The same procedure is repeated on Question 17. The significant point here is that considerably fewer people have been asked this question.

Sample size relative to full dataset	Residual Mean Square Error
95% Sample	2836.8
80% Sample	4135.37
75% Sample	5041.6
50% Sample	6501.36
30% Sample	11746.22
10% Sample	23309.9

Table 2.4: Q17: Root mean square error of sub-sample predictions

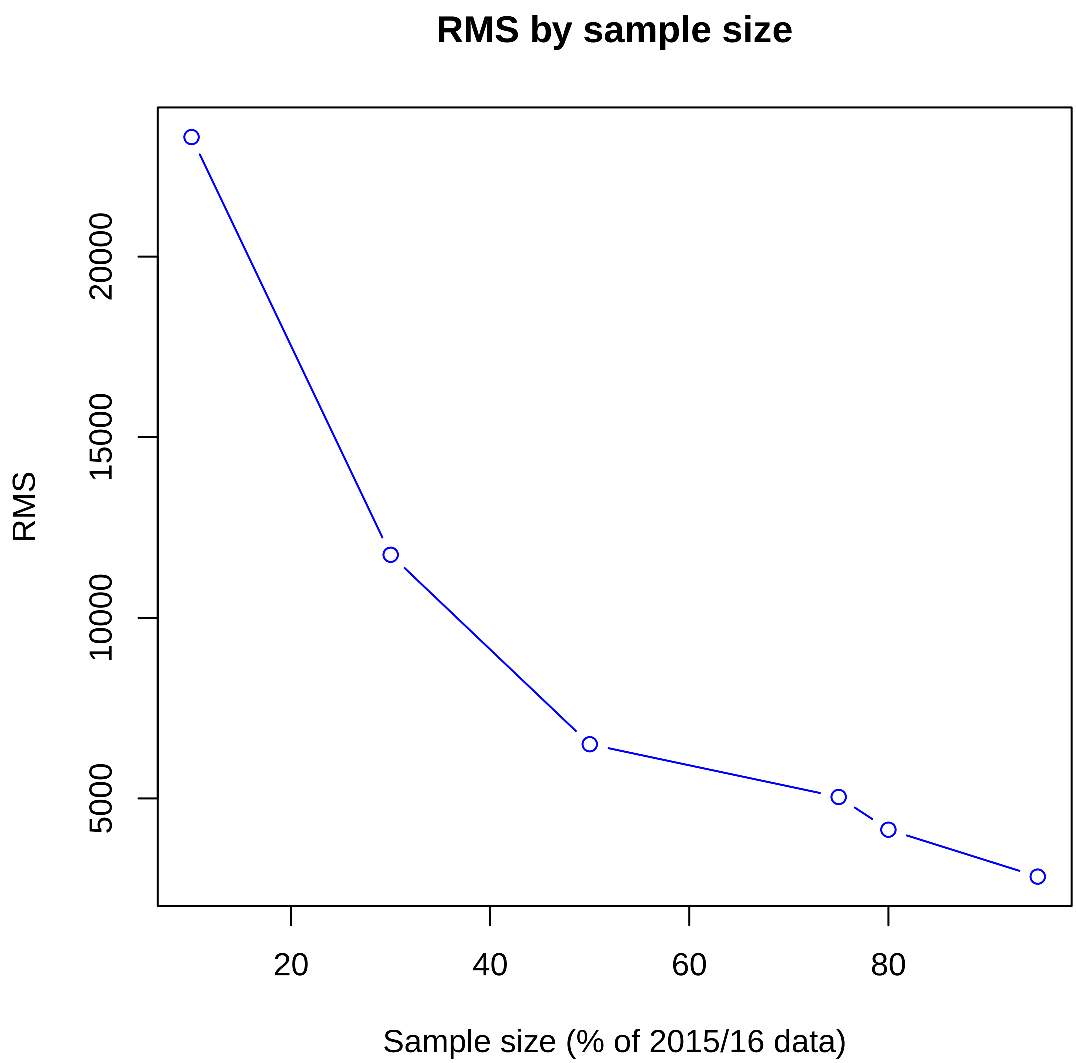

Figure 2.26: Q17 (more than monthly trips in last year): estimates Residual Mean Square for random sub-samples relative to whole dataset for 2015/16

75.Whilst full results are given in a `csv` file, figure 2.27 presents a choropleth map of the number of people in each authority area who had made a trip to the natural environment in the previous week.

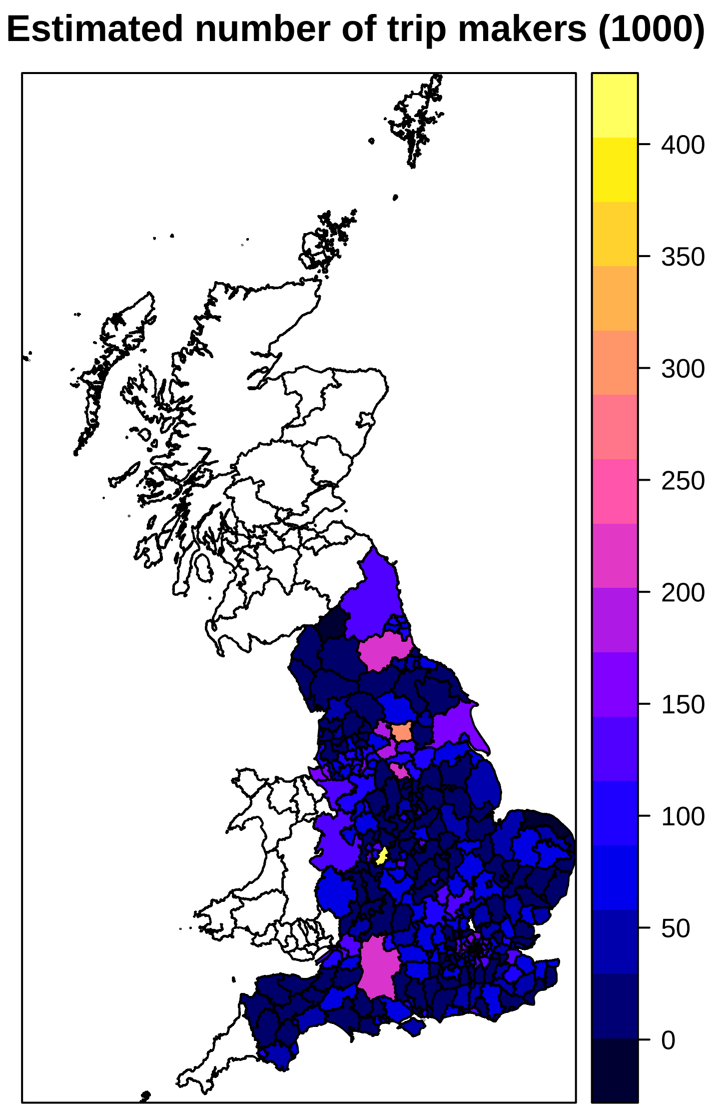

Figure 2.27: Q17: Small area estimate of the number of people in each lower level authority reporting at least one trip a month to the natural environment in the previous year based on 2015/16 MENE data

3 Conclusions

3.1 Synthetic small area estimates

76. The synthetic small area estimates for the two outcome variables (Q1 and Q17) are made available as a set of `.csv` files giving the ONS code for the small area (Local Authority and MSOA) along the posterior median and 95% credible intervals for the estimated number of people who made a trip in the last week, or who made a trip more than once of twice in the last year. The R code necessary to reproduce this analysis is available in a github repository.

77. This feasibility study has demonstrated the viability of a small area approach to the MENE survey. Tables 2.3 and 2.4 indicate a constructed Residual Mean Square Error (RMS) estimate to illustrate the likely effect of reducing sample size. It can be seen that sample size reductions have less impact in terms of Q1 (trips in the previous week) than Q17 (at least monthly trips in previous year). There are two points to note. Firstly, the RMS is computed using the entire dataset for 2015/16 as the "gold standard". Clearly, it would be preferable to have an actual "gold standard"; most methodological studies in small area estimation conduct an entirely simulation based exercise where the entire population is synthesised, a sample taken and the small area estimates compared the the simulated "gold standard" or "ground truth". Therefore these RMS estimates can not be used to compare this method with any other; they can only be used internally, to demonstrate how the quality of the estimates decreases with sample size.

78. It is apparent from several of the caterpillar plots presented that, when using data from 2015/16, the interval estimates for several local authorities are extremely wide. This is because there were few or no respondents from those local authorities in that year. There are two solutions to this problem. The first is to establish an upper level model for the survey geography. Allowing spatially correlated random effects would mean that the random effect for any local authority would be based on its neighbours values as well as its own. A more data driven approach would be to develop an upper level model based on local authority specific predictor variables. We would further recommend for any subsequent small area estimation that all available years' data should be used, which would allow a data driven model to use patterns between individuals and areas to be established, even allowing for changes over time.

79. There are some data challenges to be reconciled. Firstly, ownership of a car is a strong predictor of engagement with the natural environment, yet does not seem to be recorded in the most recent two years of the available survey. This reduces the precision by which small area estimation methods can predict a response. We would recommend that this question be restored in subsequent years. It is possible to incorporate missing data imputation methods within a Bayesian framework and this could be done in a definitive study and data from 2016/17 and 2017/2018 could be used to get stronger evidence on the relationship between other person type variables and geographical information. However, whilst Bayesian methods are powerful, they are not a magic wand and data would always be preferred to any compensatory technique.

80. There are other questions (general health) which are also only available for certain years. More significantly, there are other survey questions which have drifted from the census definitions; for example limiting long term illness in the census now records at three levels but the MENE survey has a simple yes/no question in response to disability. Small area estimation, and many techniques around non-random sampling which aim to weight a sample after the event (e.g. post-stratification) require that auxiliary variables are available which match the variables in the survey of interest. It is always a balance between maintaining a consistent data collection over time and keeping up to date with external changes. However, we would suggest that a modelling based approach allows any changes to be modelled and hence the advantages seem to lie more with consistency between the MENE survey and external data. If the MENE survey were ever to be based on a non-random sampling frame (such as happens in web-surveys) we would suggest that the need to match respondent information with auxiliary data becomes even more important.

81. It would seem that, using appropriate small area estimation methods, a reduction in sample size is possible. The challenge in this work is that it assumes that the model is essentially correct. It assumes that specific person types generally behave the same wherever they are in the country with the exception of an overall local effect captured by the local authority specific random effects. This is rarely a popular assumption with local authorities being judged on the output of such a modelling exercise. It is possible with more judicious modelling (such as spatial random effects models using the postcode sector of the respondent) to better estimate local authority specific random effects. Clearly this requires both further technical work and further discussions with stakeholders as to the likely acceptability of such modelling. Similarly, producing more granular geographical results such as MSOA level small area estimates, while feasible perhaps merits further work looking at local specific effects.

82. It is possible to produce small area estimates for finer domains, such as MSOAs. These results are given in a csv file. These may be useful in terms of identifying groups of people who are not making the most of contact with the natural environment. We would however suggest that this does not necessarily require small area estimates, but that the same understanding could be gained by carefully examining the model results. For example, the "fixed effects" provide information on types of people who have differential access to a natural environment. Local authority random effects provide information on areas of the country where access might be higher or lower than the average, assuming every part of the country had the same mix of people types.

4 Appendix

4.1 Technical clarifications

4.1.1 Odds and odds ratios

The proportion is a well recognised summary metric:

$$Proportion = \frac{\text{Number of individuals responding with 'yes'}}{\text{Total number of individuals surveyed}}$$

In many fields (including medicine) it is common to use odds as a summary metric:

$$Odds = \frac{\text{Number of individuals responding with 'yes'}}{\text{Number of individuals responding with 'no'}}$$

The odds ratio is a less commonly used summary measure of treatment / exposure effect. It is defined as:

$$Odds\ ratio = \frac{\text{Odds of 'yes' for group A}}{\text{Odds of 'yes' for group B}}$$

An odds ratio with value of (almost) 0 indicates that the odds of replying 'yes' for group 'A' are either very low (almost zero respondents replied 'yes') or the odds of replying 'yes' for group 'B' are very high (almost every respondent replied 'yes') or both. An odds ratio of 1 indicates that the odds of replying 'yes' are the same in both groups, regardless of the relative number in each group who replied 'yes'. Odds ratios above 1 indicate that relatively more respondents in group 'A' replied yes than did respondents in group 'B'.

The log-odds ratio is the (natural) logarithm of the odds ratio.

$$\textit{LogOdds ratio} = \log_e \left(\frac{\text{Odds of 'yes' for group A}}{\text{Odds of 'yes' for group B}} \right)$$

The effect of using a log-odds ratio might be for visual or interpretative reasons. An odds ratio of 0.5 and an odds ratio of 2 indicate that the odds of replying 'yes' in group 'A' are either half of twice the odds of group 'B'. There is a symmetry here which cannot be seen if these effects were to be plotted. Another reason, is that standard modelling approaches naturally work with the log-odds ratio for their own reasons.

As a point of clarification for later, it is possible to convert a proportion into an odds using formula 4.1

$$Odds = \frac{Proportion}{1 - Proportion} \tag{4.1}$$

4.1.2 Logistic regression models

This analysis considers two dichotomous outputs, namely whether a respondent indicated that they have made at least one trip in the last week, or whether over the course of the previous year they had made more than one trip a month. Denoting the number of individuals by n, and a single individual with a subscript i (where $i = 1, \ldots, n$). for individual i, the response variables are created as:

Question 1

$$Y_i = \begin{cases} 1 & \text{if} \quad \text{if one or more trips have been made in the last week} \\ 0 & \quad\quad \text{otherwise} \end{cases}$$

Question 17

$$Y_i = \begin{cases} 1 & \text{if} \quad \text{if trips have been made more than once a month in the last year} \\ 0 & \quad\quad \text{otherwise} \end{cases}$$

For technical reasons, the standard statistical method for fitting a model to such data is the so-called logistic regression. There are three elements to these models.

1. Assume that the response variable Y_i is a realisation of a Bernoulli random variable where the probability $Y = 1$ is given by θ_i.

2. Next assume that there it is resonable to relate the outcome to a standard linear predictor of the form $\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ where β_j (with the subscript $j = 1, \ldots, p$ denoting a distinct predictor variable from 1 to p) are the parameters to be estimated and $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})$ are the values of the predictor variables for individual i.

3. In many ways, the η_i values are similar to the $\hat{y}$ values of a standard linear regression. However, we need a way of mapping these linear predictor values onto the probability θ_i that individual i reported having taken trip in the last week, or a trip more than once a month in the last year. To do this, a so-called logistic link function is used.

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \eta_i \qquad (4.2)$$

Equation 4.2 is referred to as the *logistic link* function and is the reason these models are referred to as “logistic regression” models. It is something of a default in medical research where they have been used for over a century. In 1972 [1] developed a flexible framework for a range of related models. From the point of view of someone wishing to interpret the parameter estimates from a logistic regression model, it can be seen that there is a similarity between this function and the formula given in equation 4.1. What this means is that the entire linear predictor (and the intercept on its own) are effectively the (natural) log odds of a respondent answering a question with ‘yes’. The other parameters are the (natural) log-odds ratio for a ’yes’ response for a member of the group described by that predictor variable relative to the baseline. For example, the log-odds ratio of a male respondent relative to a female replying ‘yes’ to the question about trips.

[1] Nelder, J. A. and R. W. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General) 135*(3), 370–384

For reference purposes, the inverse function to the logistic link 4.2 is as follows:

$$\theta_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$$

In other words, we can use this to calculate the probability that a particular respondent replied ‘yes’ to the dichotomised trip question.

5 Bibliography

Nelder, J. A. and R. W. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General) 135*(3), 370–384.