

Small Area Income Estimates: Model-Based Estimates of the Mean Household Weekly Income for Middle Layer Super Output Areas, 2013/14 Technical Report

Office for National Statistics

© Crown Copyright 2016

December 2016

Official Statistics

ONS official statistics are produced to the high professional standards set out in the Code of Practice for Official Statistics.

About us

The Office for National Statistics

The Office for National Statistics (ONS) is the executive office of the UK Statistics Authority, a non-ministerial department which reports directly to Parliament. ONS is the UK government's single largest statistical producer. It compiles information about the UK's society and economy, and provides the evidence-base for policy and decision-making, the allocation of resources, and public accountability. The Director-General of ONS reports directly to the National Statistician who is the Authority's Chief Executive and the Head of the Government Statistical Service.

The Government Statistical Service

The Government Statistical Service (GSS) is a network of professional statisticians and their staff operating both within the Office for National Statistics and across more than 30 other government departments and agencies.

Contacts

This publication

For information about the content of this publication, contact Nigel Henretty
Tel: +44 (0)1329 44 7934
Email: better.info@ons.gsi.gov.uk

Other customer enquiries

ONS Customer Contact Centre
Tel: 0845 601 3034
International: +44 (0)845 601 3034
Minicom: 01633 815044
Email: info@statistics.gsi.gov.uk
Fax: 01633 652747
Post: Room 1.101, Government Buildings,
Cardiff Road, Newport, South Wales NP10 8XG
www.ons.gov.uk

Media enquiries

Tel: 0845 604 1858
Email: press.office@ons.gsi.gov.uk

Copyright and reproduction

© Crown copyright 2016

You may re-use this information (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence.

To view this licence, go to:

www.nationalarchives.gov.uk/doc/open-government-licence/

or write to the Information Policy Team, The National Archives, Kew, London TW9 4DU

email: psi@nationalarchives.gsi.gov.uk

Any enquiries regarding this publication should be sent to: info@statistics.gsi.gov.uk

This publication is available for download at: www.ons.gov.uk

Table of Contents

Executive Summary	5
1. Introduction	6
2. Background to the Need for Income Estimates for Small Areas	7
2.1 Income Question in the 2001 Census	7
2.2 Department for Work and Pensions Benefit Data	8
2.3 Small Area Estimation and Modelling	8
3. Methodology	9
4. Modelling For Income	9
4.1 Introduction	9
4.2 The Data Sets	10
4.2.1 Survey Data.....	10
4.2.2 Covariate Data Sets.....	13
4.3 Developing the models.....	14
4.3.1 Total Weekly Household Income (unequalised).....	15
4.3.2 Net Weekly Household Income (unequalised)	19
4.3.3 Net Weekly Household Income – Equalised, Before Housing Costs.....	22
4.3.4 Net Weekly Household Income – Equalised, After Housing Costs	26
4.3.5 Observations.....	29
5. Results of Modelling for Income	30
5.1 Total Weekly Household Income (unequalised)	30
5.2 Summary of Results.....	32
6. Quality of the Estimates	33
6.1 Residual vs. Model Estimates Diagnostic Plot.....	33
6.2 Model vs. Sample Estimates Diagnostic Plot.....	35
6.3 Coverage Diagnostic.....	37
6.4 Wald Statistic	38
6.5 Stability Analysis	38
6.6 Diagnostic Results	40
6.7 Conclusions	41
7. Comparing results for 2011/12 and 2013/14, and measuring change	41
7.1 Models	41
7.2 Diagnostics	41
7.3 Estimates	42
7.3.1 Covariates	42

7.3.2 Geography of estimation.....	43
7.4 Estimates of change.....	43
8. Guidance on the Use of the Estimates.....	43
Appendix	45
A. Model Procedures	45
A.1 Basic Theory	45
A.2 Basic Theory	45
A.3 General SAEP Theory.....	46
A.4 Small Area Estimation (SAEP) Income Model.....	47
A.5 Adding auxiliary data to the model	48
A.6 Benchmarking.....	48
B Data Sources.....	49
B.1 Survey Data Income - Definitions	49
B.2 Total household weekly income (unequalised).....	49
B.3 Net household weekly income (unequalised)	50
B.4 Net household weekly income before housing costs (equalised).....	50
B.5 Net household weekly income after housing costs (equalised).....	50
B.6 FRS and Households Below Average Income (HBAI) Data	50
B.7 Auxiliary Data Sources and Covariates	51
C Data Preparation	58
D Results of Modelling for Income	58
E Diagnostic Results.....	65
F Calculation of Direct Survey Estimates and Confidence Intervals	79
G Bibliography	82

Executive Summary

In order for government, local authorities and other bodies to identify areas of poverty, data at the smallest possible geographical level are required. For a number of reasons it was not considered appropriate to include a question on income in the 2011 Census, an alternative approach has been to combine survey data with information from other sources through the use of small area estimation methods.

This report provides technical information about the methods and processes used to produce the Middle-layer Super Output Area (MSOA) estimates of average household income for 2013/14. It follows the previous publication of MSOA income estimates, for 2011/12. Estimates are produced for the following four income types:

- total household weekly income (unequalised);
- net household weekly income (unequalised);
- net household weekly income before housing costs (equalised); and
- net household weekly income after housing costs (equalised).

Results for England and Wales show that higher levels of income are found in the South of England particularly around London. The South West and North East of England, and Wales show lower income levels.

A number of diagnostic checks are used to assess the model fit and quality of the estimates. The checks show that in general the models are well specified and the modelling assumptions are satisfied. This provides assurance of the accuracy of the estimates and the confidence intervals produced from the models.

Comparisons of the estimates over time should be made with caution. They represent the mean weekly household income for the reference time period, but are not optimised to give a measure of change. Section 7.4 provides further guidance about appropriate use of the estimates for identifying change over time.

It should be noted that these model based MSOA estimates of average household income are not calculated in the same way as the national and regional household income data published separately by ONS. As well as the output geography, the definitions of income and data sources employed are different.

1. Introduction

There is a specific and increasing interest from government, local authorities and many other bodies in obtaining income data at the smallest possible geographical level. This information is needed in order to help identify deprived and disadvantaged communities and to support work on social exclusion and inequalities. The requirement for data on income was previously reflected by Census User Groups who made a strong case for a question on income to be included in the 2001 Census. Although this need was recognised by the government, concerns were raised about public acceptance and the risks to the overall Census returns. As a result a question on income was not included in either the 2001 or 2011 Census. Alternative methods for obtaining data on income at the small area level were identified and implemented. One of the options identified was the use of small area estimation methodologies to produce small area income estimates.

The method for producing small area estimates combines survey data with auxiliary data that are correlated with the target variable. The approach is to create a model which relates the survey variable of interest (e.g. income) to these auxiliary variables (covariates). The survey sample is too small to provide reliable direct estimates for small areas or domains but synthetic estimates can be made based upon the model parameters and values for the covariate data which are available for all of the small areas. These estimates and confidence intervals were originally released as experimental statistics¹ on the ONS Neighbourhood Statistics website in 2005 and are now classified as National Statistics.

A requirement for estimates of average weekly household income by Middle Layer Super Output Area (MSOA) was identified. Super Output Areas (SOAs) are a geographic hierarchy designed to improve the reporting of small area statistics in England and Wales. A range of areas have been developed that are of consistent size and are subject to minimal boundary changes. These areas are built from groups of Output Areas (OAs) used for the 2011 Census. The SOA layers form a hierarchy based on aggregations of OAs, these add firstly to form Lower Layer Super Output Areas (LSOA) then to larger areas. MSOAs have a mean population of 7,200 and a minimum population of 5,000. They are built from groups of LSOAs and constrained by the local authority boundaries used for 2011 Census outputs.

This report is a technical guide to support the 2013/14 set of MSOA level income estimates for England and Wales. Chapter 2 provides background information including a detailed description of the requirement for estimates of household income at the small area level. Chapters 3 and 4 describe the methods used and their application for estimating average household income at MSOA level.

Chapter 5 shows the modelled estimates with their respective confidence intervals, an assessment of the quality of the estimates is demonstrated using diagnostic plots in Chapter 6. Chapter 7 shows a comparison of the model (and covariate data) used to derive the income estimates for 2013/14 with that used for 2011/12 and discusses use of the estimates for measuring change in average income. Finally chapter 8 provides further guidance on use of the 2013/14 estimates in practice. All technical details of the methodology are contained in the appendices.

¹Experimental statistics are in the testing phase and are not yet fully developed. A guide to Experimental Statistics is available at : <https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/guidetoexperimentalstatistics>.

2. Background to the Need for Income Estimates for Small Areas

There is a specific and increasing interest in obtaining income data at the smallest possible geographical level. Interest stems from a variety of sources: central government departments, local authorities, academics, commercial organisations and independent researchers. These data are essential for the identification of deprived and disadvantaged communities, evaluation research, provision of information for practitioners, and for the profiling of geographical areas.

This need was reflected in the White Paper on the 2001 Census of Population that included the following statement:

"Consultations with users throughout 1995-98 have indicated a widespread requirement to have information on the level of individual gross income available from the Census. Income is widely seen as a more discriminating variable than occupation or housing condition for the purposes of identifying areas of affluence or deprivation and in economic and social research. All main user groups made a strong case for a question on income in their business cases for census topics. In particular, central and local government users expressed a requirement for the information to be used to support a range of activities including resource allocation, policy and development review, the derivation of deprivation indicators, and in the assessment of inequalities and social exclusion".

Although this need was recognised by the government, concerns were also expressed about user acceptance to a question on income and the risks to the overall response to the Census. As a result the Government Statistical Service set up a working group (the Income Data Working Group) to investigate the feasibility of meeting users' requirements for income data from alternative sources. The report produced by the working group combined the results of this research with an outline of users' requirements gathered as part of the consultation on requirements for the Census. The report provided an overview of the strengths and weaknesses of three key options:

- Including an income question in the 2001 Census of Population
- Using benefit data from the Department Work and Pensions (DWP)
- Developing small area estimation/modelling techniques

2.1 Income Question in the 2001 Census

The Income Data Working Group found that most users (59%) had a preference for including an income question in the 2001 Census of Population. However, a number of inner city authorities were more concerned about the acceptability of the income question and the implications for response, and consequently did not favour including an income question in the Census.

The Census Offices conducted a series of tests with the overall objective of identifying the most effective method for collecting information on income from a self-completion questionnaire. The 1997 Census Test showed that a question on income:

- Failed to elicit accurate information;
- attracted the highest number of objections of any Census question;
- lowered overall response rates significantly; and

- was not answered by a relatively high percentage of respondents (Teague (1999)).

The report, 'Income Data for Small Areas', was circulated to users and their comments were taken into account when the Government decided in January 2000 not to include a question on income in the 2001 Census. The same decision was taken for the 2011 Census. In 2015, ONS carried out a consultation on topics to be included in the 2021 census questionnaire. The results of this consultation will help determine whether a question on household income will be asked in the 2021 Census.

2.2 Department for Work and Pensions Benefit Data

The Income Data Working Group examined the feasibility of using data from the Department for Work and Pensions (DWP) on the receipt of benefits in order to provide information on income. Of most value would be Income-Related Benefits (IRBs) which are only payable if the recipient's income is below a certain threshold. These data are from an administrative source, thus they are not subject to sampling error. In addition, the data are frequent and timely and can be analysed by a range of other variables (e.g., age, sex, length of claim) and can be analysed by different geographies (as individual cases are postcoded).

DWP data are a rich source of information on the relative incidence of low incomes. The data however, have disadvantages:

- IRB entitlement cannot be equated with low income; recipients of in work IRBs tend to have higher incomes than those on Income Support or Jobseekers Allowance. Entitlement can also be linked to savings, so a person with savings may not receive a benefit even if they have a low income
- Benefit levels are not constant. Therefore, changes cannot be assumed to match changes in poverty thresholds
- IRB take-up varies between groups (e.g., it can be very high for lone parents but lower for pensioners)
- IRB take-up can vary between areas
- The eligibility criteria for the IRBs inevitably means that some groups in poverty will not be captured (e.g., in-work poor)
- IRBs provide little or no information for the middle and higher ends of the income distribution

2.3 Small Area Estimation and Modelling

Small area estimation is used to improve the precision of survey estimates for small areas or domains. Surveys are designed to provide reliable estimates at national and sometimes regional levels but are not typically designed to provide estimates at small area level (e.g. local authorities, output areas, etc.). With the exception of the Labour Force Survey all the principal national household surveys have a clustered design. This means that the sample is not distributed totally randomly across England and Wales, but that certain areas are first selected as primary sampling units (PSUs) and then households are selected for interview from these. The areas selected as PSUs are postcode sectors. The selection of postcode sectors is stratified in such a way that their distribution is nationally representative. The problem for deriving direct survey estimates at small area level is that, irrespective of the total sample size, with a clustered sample design a large proportion of areas (such as MSOAs) contain no sample respondents at all and so direct estimate would not be possible.

Also, where there is a sample for particular MSOAs, the sample sizes are likely to be so small that the variability around the estimates would be too high for reliable estimates.

Following some preliminary studies into small area estimation, Methodology Directorate of the Office for National Statistics established the Small Area Estimation Programme (SAEP) in April 1998. The overall aim of the SAEP was to establish statistical methodologies for deriving estimates from variables contained in social surveys, for areas defined by a variety of boundary systems and that account for the clustered sample design.

3. Methodology

The technique of synthetic estimation produces estimates for domains, in which survey data are insufficient, by borrowing strength from other data sources. The other data sources (known as auxiliary data or covariates) are available on an area basis and for all areas in the target population. At the level of these small areas, sample survey sources are not generally available so the covariate data are usually from some administrative system or from a previous census.

The small area estimate is based on the area level relationship between the survey variables and auxiliary variables. This relationship can be fitted by regressing individual survey responses (e.g. weekly household income) on area level values of the covariates (e.g. proportion of MSOA population claiming Income Support). The fitted model describes the relationship between the area level summary (mean) values of the target survey variable and the covariates.

While the model has been constructed only on responses from sampled areas, the relationships identified by the model are assumed to apply nationally. Thus as administrative and census covariates are known for all areas, not just those sampled, the fitted model can be used to obtain estimates and confidence intervals for all areas. This is the basis of the synthetic estimation that ONS has used in its development of small area estimation. For more technical details of the SAEP methodology see Appendix A.

Once a model has been selected an assessment of the quality is made using a number of diagnostics; these are detailed in Chapter 6.

4. Modelling For Income

4.1 Introduction

This chapter describes how the general SAEP methodology has been used for estimating average household income at the MSOA level. The data sets (both survey and covariate) used in the modelling process are described as well as the final models. The estimates obtained from the models are also displayed.

4.2 The Data Sets

4.2.1 Survey Data

The survey data were obtained from the 2013/14 Family Resources Survey (FRS). The FRS was chosen as the source for survey data for this study since it is the survey with the largest sample that includes suitable questions on income. The Labour Force Survey (LFS) also includes questions on income but was not used because it did not cover the full target population and does not record all sources of income (i.e. it measures income for employees only and no account is taken of the self-employed, income from benefits or housing costs).

The FRS allows four survey variables to be modelled and the average is used as the summary variable, i.e. the estimates produced are values of average MSOA income for the following four income types:

- Total household weekly income (unequalised)
- Net household weekly income (unequalised)
- Net household weekly income before housing costs (equalised)
- Net household weekly income after housing costs (equalised)

Equalised income means that the household income values have been adjusted to take into consideration the household size and composition; it represents the income level of every individual in the household. Equalisation is needed to make sensible income comparisons between households. For more details on these income definitions see Appendix B.

These estimates use the OECD equalisation scale. This was in response to the Government's 2004 Spending Review, which stated that future child poverty measurements will report incomes before housing costs and equalised using the OECD scale. More information on the equalisation scale is available in Appendix B.

The FRS uses a stratified clustered probability sample drawn from the Royal Mail's small users Postcode Address File (PAF). The survey selects 1,417 postcode sectors with a probability of selection that is proportional to size. Each sector is known as a Primary Sampling Unit (PSU). Within each PSU a sample of addresses is selected. In 2013/14, 24 addresses were selected per PSU. More information on the FRS methodology is contained within the FRS technical report (Shale et al (2015)).

The FRS aims to interview all adults in a selected household. A household is defined as fully co-operating when it meets this requirement. In addition, to count as fully co-operating, there must be less than 13 'don't know' or 'refusal' answers to monetary amount questions in the benefit unit schedule (i.e. excluding the assets section of the questionnaire). In 2013/14 the achieved sample size (for the UK) was 20,142 households.

The requirement for this project is to produce MSOA level estimates of average household income (four types) for England and Wales. The survey data file used contained 15,177 households from 1,173 postcode sectors. The final survey data file for England and Wales contained cases in 2,547 different MSOAs out of a total of 7,201. The number of cases per MSOA in the achieved FRS sample varies widely particularly due to the fact that MSOAs cut

across the postcode sectors primary sampling unit. For example, some MSOAs recorded only 1 response whereas, others had 32 (the maximum number of sampled households).

For each different income type a number of records were found with values of income less than or equal to £1, these were removed from the sample data set. Additional records with extremely high total income values were removed as they would have had an unduly large influence on the model². For the net weekly (unequalised and equalised) income, records were removed where the net income was greater than the total income by £10. The net equalised weekly income excludes households containing a married adult whose spouse is temporarily absent. This is because the data for net weekly income come from another Family Resources Survey dataset, called the Households Below Average Income data³.

The final sample sizes for each income type for England and Wales are recorded in Table 1.

Table 1: Survey sample sizes for income types, England and Wales

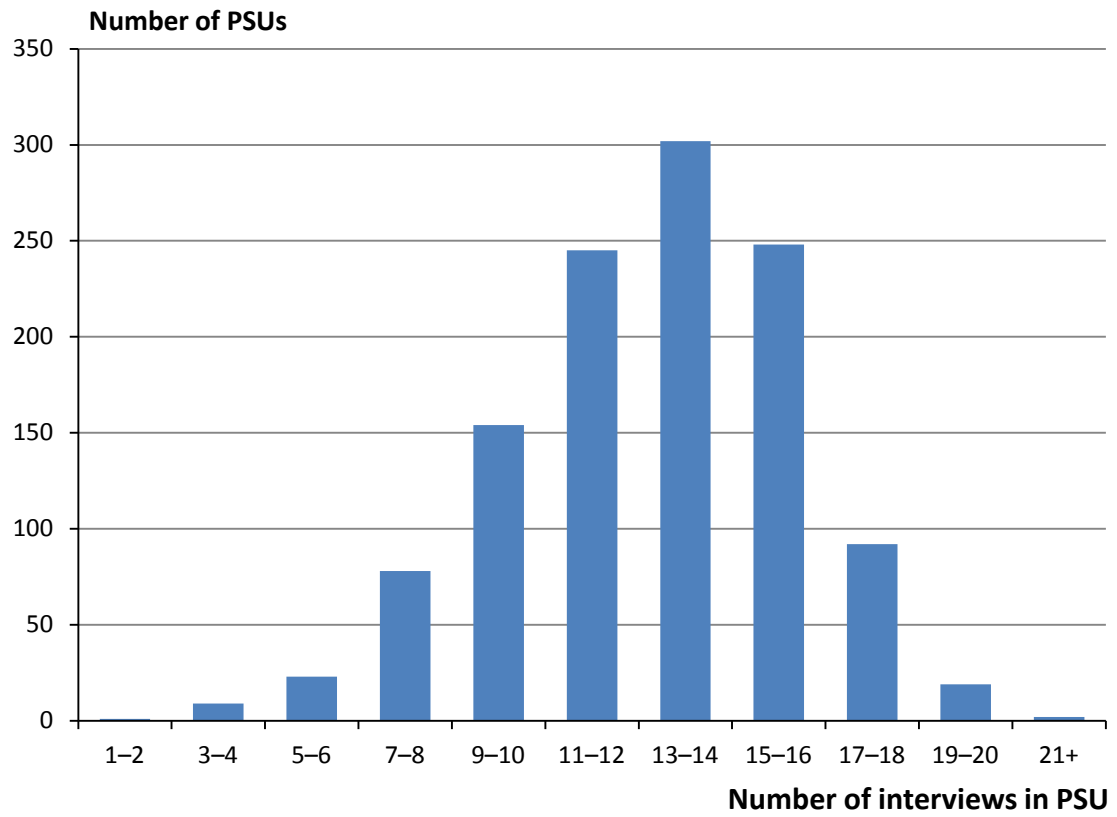
Weekly household income type	No. & % of households removed	No. households in final sample	No. postcode sectors in sample	No. msoas in sample
Total (unequiv)	186(1.23%)	14,991	1,173	2,539
Net (unequiv)	310(2.04%)	14,867	1,173	2,535
Net (equiv) before housing costs	309(2.04%)	14,868	1,173	2,535
Net (equiv) after housing costs	493(3.25%)	14,684	1,173	2,530

Figure 1 displays the distribution of interviews by primary sampling unit. The FRS has a response rate of around 62%. Since 24 addresses were selected by PSU, one would expect, as shown, the number of PSUs to peak at around 14.

²These households either had a total weekly household income which equated to over £1,000,000 per year, or a total weekly household income over £15,000 and were the only household sampled in a MSA.

³ The Households Below Average Income dataset is an unpublished record level dataset maintained by the Department for Work and Pensions. More information about it is available from the data.gov website.

Figure 1: The number of interviews achieved in each Postcode Sector, England and Wales, FRS



4.2.2 Covariate Data Sets

The SAEP methodology requires covariate data to be available at a geographic level compatible with MSOAs. A range of data sources were used in the modelling process that were considered to be related to household income. In all cases the sources provided are related to household income. They are:

- Census, 2011
- Department for Work and Pensions benefit claimant counts, August 2013
- Valuation Office Agency Council Tax Bandings, March 2013
- Her Majesty's Revenue and Customs, Child Tax Credit and Working Tax Credit, Aug 2013
- Office for National Statistics, House Price Statistics for Small Areas, Q1 2014
- Department of Energy & Climate Change, Energy Consumption data, 2013
- Regional/country identification variable

The covariates used for modelling income were the same for England and Wales with the exception of the Council Tax Banding data. Council Tax bands are available for both England and Wales on the Neighbourhood Statistics website; however, the values of the bands are defined differently. For this reason the Council Tax covariates in the models appear separately for England and Wales. For more information on the Council Tax bands see Appendix B.

The data used are as close to the reference time period of the target income estimates as possible (i.e. for 2013/14). Administrative data are collected primarily for government administrative processes and may change over time. The DWP data sources for benefit claimants and HMRC data sources for Tax Credits have changed since the reference time period of these estimates. More information about the variables considered for inclusion in the model and the recent changes to the sources is provided in Appendix B.

4.3 Developing the models

Linear models that take into account the fact that each individual household belongs to a specific area were developed for England and Wales. These models take the survey variable, weekly household income, as the response variable and the area level covariates as explanatory variables. The models relate the survey variable of interest (measured at household level) to the covariates that relate to the small area in which the household is located. Once fitted the models can be used to produce estimates of the target variable at the small area level, i.e. the models can be used to produce MSOA level estimates of average household weekly income and calculate confidence intervals for the estimates.

For all four types of income the response variable, weekly household income, is not normally distributed but positively skewed (the largest values differ from the mean more than the smaller values do). By using the natural logarithm (ln) of the appropriate type of income as the response variable this skewness is reduced and it is assumed for the analysis that the transformed variable follows a normal distribution.

The models were fitted using the statistical software SAS with postcode sectors at the higher level and households at the lower level. Region/country indicator terms are forced into the model (whether significant or not) and then the method of step-wise forward selection (see Appendix A) is used to identify the significant covariates to be included in the models from the set of covariates given in Appendix B.

All of the appropriate covariates (those expressed as percentages or proportions) were transformed onto the logit scale and both the transformed and original covariates were considered for inclusion in the models. The covariates were centred by subtracting the corresponding means for England and Wales. Centring the covariates enables easier interpretation of the model parameters, e.g. the intercept now represents the weighted average of the response variable (after the ln transformation) over all areas.

Initially, significant covariates were selected for inclusion in the models. Then with these significant covariates, interaction terms were created, tested for significance and where appropriate included in the models.

For the 2001/02 estimates the need for separate models for England and Wales was analysed and the conclusion drawn that a single model was appropriate, this was employed for all subsequent estimates.

After modelling, adjustments are made to the modelled estimates to ensure they are consistent with the direct survey estimates at regional level for England and country level for Wales (this is known as benchmarking). The FRS survey data are used to calculate direct estimates of income at these higher geographical levels (estimates at this level are considered robust). The model-based MSOA estimates of income are aggregated to this region/country level and comparisons made between the two sets of estimates. The ratio of direct survey estimate to aggregated model estimate at the region/country level is used to scale all model MSOA-level estimates and their confidence intervals. More details on this benchmarking methodology and aspects of the modelling methodology are given in Appendix A.

The subsequent sections describe the models developed for the four income types for England & Wales.

4.3.1 Total Weekly Household Income (unequalised)

The model selected to estimate total weekly household income was:

$$\ln(y_{ij}) =$$

6.341 (0.023)	- Constant	
- 0.005(0.038) northeast _k		} Region/Country
- 0.029(0.030) northwst _k		
- 0.031(0.033) york _k		
- 0.061(0.035) eastmid _k		
- 0.120(0.040) westmid _k		
- 0.061(0.031) east _k		
- 0.014(0.028) southeast _k		
- 0.097(0.034) southwst _k		
- 0.105(0.036) wales _k		
+ 1.000(0.109) phrpman _k		
+ 0.026(0.001) ewavhhpeop _k		
+ 0.131(0.035) lnpemployd _k		
- 0.032(0.008) ewfamwkbfe _k		
- 0.080(0.030) lnpcpf _k		
+ 3.281(0.994) phhtype4 _k		
- 2.200(0.771) ewjsafemale _k		
+ 16.361(4.740) ewpdccd12years _k		
+ 0.113(0.036) ln16_59 _k		
- 1.506(0.752) ewpcp25years _k		
+ 0.021(0.010) ewfamwkw _k		} Interactions
+ 8.446(3.722) ewpdccd12years_ewfam _k		
+ 0.051(0.025) ewavhhpeop_northwst _k		
+ 4.852(2.475) phhtype4_york _k		
+ $u_j + e_{ij}$		

$$\hat{\sigma}_u^2 = 0.0035 (0.002)$$

$$\hat{\sigma}_e^2 = 0.5390 (0.006)$$

Equation [1]

Details of the various components included in the model are outlined below. The figures in parentheses are the standard errors of the coefficients and the variables have the following labels:

y_{ij} = weekly income of household i in postcode sector j ;

u_j = area level random residual for postcode sector j ;

e_{ij} = within area residual for household i in postcode sector j ;

$\hat{\sigma}_u^2$ = estimated variance of u_j ;

$\hat{\sigma}_e^2$ = estimated variance of e_{ij} ; and

the subscript k relates to the MSOA that household i in postcode sector j falls within.

Table 2 contains a key to the labels of the covariates. The covariates have been grouped by source. Appendix A contains more information on the form of the model.

Table 2: Key to covariates included in the model for total household weekly income, unequivalised

Covariate Name	Label	Source	T ratio = $\left(\frac{\beta}{s.e} \right)$
northeast	North East	Country/regional indicators	-0.13
northwst	North West	Country/regional indicators	-0.96
york	Yorkshire and The Humber	Country/regional indicators	-0.94
eastmid	East Midlands	Country/regional indicators	-1.72
westmid	West Midlands	Country/regional indicators	-3.03
east	East of England	Country/regional indicators	-1.97
southeast	South East	Country/regional indicators	-0.48
southwst	South West	Country/regional indicators	-2.89
wales	Wales	Country/regional indicators	-2.95
phrpman	Proportion of HRPs aged 16 to 74 whose NS-SEC is 'managerial and professional'	Census	9.15
ewavhhpeop	Average number of people per household	Census	2.70
lnpemployd	Logit of Proportion of people aged 16 to 74 who are employed or self-employed	Census	3.79
ewfamwkbfe	Families in Work Receiving; from the Childcare Element	HMRC	-4.19
lnpcpf	Logit of Proportion of females aged 60 and over claiming Pension Credit	DWP	-2.65
phhptype4	Proportion of households that are lone parent with all child(ren) non - dependent	Census	3.30
ewjsafemale	Proportion of females aged 16 and over claiming Job Seekers Allowance	DWP	-2.85

ewpdccd12years	Proportion of people claiming Disability Living Allowance with a claim duration of 1-2 years	DWP	3.45
lnp16_59	Logit of Proportion of people aged 16 to 59		3.12
ewpcp25years	Proportion of people aged 60 and over claiming Pension Credit, with a claim Duration of 2-5 Years	DWP	-2.00
ewfamwkwt	Families in Work Receiving; Working Tax Credit Only	HMRC	2.03
ewpdccd12years_ewfamwkbfe	Interaction between ewpdccd12years and ewfamwkbfe		2.27
ewavhhpeop_northwst	Interaction between ewavhhpeop and northwst		2.00
phh4type4_york	Interaction between phh4type4 and york		1.96

With no covariates included in the model the estimated standard residual area variance $\hat{\sigma}_u^2$ is 0.0435 (0.0037) compared with 0.0035 (0.002) when the significant covariates are included in the model, a decrease of 91.98%. Therefore, these covariates together account for 91.98% of the total between area variance. As one would expect, since the covariates are at the MSOA-level, the model explains a lot of the between area variance but does not significantly reduce the within area variance. This indicates that the model is only appropriate for estimating at the area or MSOA-level.

Note some covariates have been included in the model even though they are not considered to be significant using the T rule, as described in Appendix A, since they are included in an interaction term which is significant.

The most significant covariate in the model is the Census covariate ‘phrman’, which has a T value of 9.15. As one would expect this covariate has a positive coefficient; as the proportion of the MSOA household reference persons aged 16 to 74 whose NS-SEC is managerial and professional increases so does the average weekly household income for that MSOA. ‘Inpemployd’ is the next most significant covariate in the model with a positive coefficient, and has a T-value of 3.79. It shows that as the proportion of people in an MSOA aged 16 to 74 who are employed or self-employed increases so does the average weekly household income. The relationship of a covariate with the average weekly household income may be different if it is also involved in a model interaction. For example, ‘ewavhhpeop’ is included in a model interaction with ‘NorthWst’. This suggests that the relationship between ‘ewavhhpeop’ and the average weekly household income is different for North West MSOAs.

The most significant variable with a negative coefficient (-0.032) was ‘ewfamwkbfe’ which refers to families in work receiving childcare element of working tax credit, and has a T value of -4.19. This means that a lower proportion of families claiming this benefit within an area is associated with a higher average weekly household income.

The MSOA estimates of total weekly household income obtained from the model are combined with estimates of the number of households in each MSOA and aggregated to a REGION/country level (see Appendix A). These aggregated model estimates are compared with direct estimates obtained from the survey data in order to calculate benchmarking

ratios, Table 3. The ratios in Table 3 are used to adjust the model-based estimates and their confidence intervals at the MSOA-level.

Table 3: Benchmarking results for total weekly household income (unequivalised)

Country/Region	Survey estimate	Aggregated model estimate	Ratio of survey to model estimate
North East	614	664	0.92
North West	674	680	0.99
Yorkshire and the Humber	674	657	1.03
East Midlands	684	694	0.99
West Midlands	654	659	0.99
East of England	805	755	1.07
London	945	854	1.11
South East	893	853	1.05
South West	714	695	1.03
Wales	636	626	1.02


Due to concerns raised over differences between estimates for some London areas and other published estimates, for the 2013/14 release an alternative procedure was examined which grouped London into two subregions. One region comprised boroughs in North West, West, South West and South London with the other covering South East, East and North East and North London. Such a grouping meant that their sample sizes were approximately equal and of a comparable size to that of some other regions. Benchmarking was conducted for each separately and also modelling included them as separate indicator covariates.

Although the outputs resulting from the models which separate Greater London demonstrated some difference in behaviour, the results indicated it is not currently necessary to specifically require the accepted final model to stipulate a sub-London effect, and therefore the final models chosen do not separate the two London sub-regions as part of the modelling or calibrations.


4.3.2 Net Weekly Household Income (unequivalised)

The model selected to estimate net weekly household income (unequivalised) was:

$$\begin{aligned}
 \ln(y_{ij}) = & \\
 & 6.129 (0.025) \\
 & + 0.021(0.040) \text{northeast}_k \\
 & - 0.053(0.032) \text{northwest}_k \\
 & - 0.041(0.035) \text{york}_k \\
 & - 0.044(0.036) \text{eastmid}_k \\
 & - 0.086(0.032) \text{westmid}_k \\
 & - 0.051(0.031) \text{east}_k \\
 & - 0.037(0.030) \text{southeast}_k \\
 & - 0.083(0.034) \text{southwest}_k \\
 & - 0.091(0.039) \text{wales}_k \\
 & + 0.042(0.012) \text{ewavhhpeop}_k \\
 & + 0.730(0.109) \text{phrpman}_k \\
 & + 0.118(0.036) \text{lnphhtype4}_k \\
 & + 0.106(0.038) \text{lnphhtype5}_k \\
 & - 0.146(0.043) \text{lnpltli}_k \\
 & + 19.487(4.474) \text{ewpdccd12years}_k \\
 & - 1.975(0.719) \text{ewjsafemale}_k \\
 & + 9.426(4.003) \text{ewpdlaman}_k \\
 & - 0.038(0.017) \text{lnispfemale}_k \\
 & - 1.378(0.694) \text{ewpcp25years}_k \\
 & + 0.268(0.074) \text{pflat}_k \\
 & + 0.067(0.033) \text{lnphhtype7}_k \\
 & - 0.138(0.054) \text{lnpltli_southeast}_k \\
 & + 0.052(0.021) \text{lnpltli_ewavhhpeop}_k \\
 & - 27.957(11.932) \text{ewpdccd12years_south}_k \\
 & - 41.226(16.682) \text{ewpdccd12years_wales}_k \\
 & + 15.553(6.707) \text{ewpdlaman_lnphhtype4}_k \\
 & + u_j + e_{ij}
 \end{aligned}$$



Regional/Country



Interactions

$$\begin{aligned}
 \hat{\sigma}_u^2 &= 0.001(0.002) \\
 \hat{\sigma}_e^2 &= 0.435(0.005)
 \end{aligned}$$

Equation [2]

Variable labels are as in Equation [1].

Table 4 contains a key to the labels of the covariates.

Table 4: Key to covariates included in the model for net household weekly income (unequivalised)

Covariate Name	Label	Source	T ratio = $\left(\frac{\beta}{s.e}\right)$
Northeast	North East	Country/regional indicators	0.52
Northwest	North West	Country/regional indicators	-1.65
York	Yorkshire and The Humber	Country/regional indicators	-1.17
eastmid	East Midlands	Country/regional indicators	-1.23
westmid	West Midlands	Country/regional indicators	-2.68
East	East of England	Country/regional indicators	-1.65
southeast	South East	Country/regional indicators	-1.23
southwst	South West	Country/regional indicators	-2.46
wales	Wales	Country/regional indicators	-2.30
ewavhhpeop	Average number of people per household	Census	3.58
phrpman	Proportion of HRPs aged 16 to 74 whose NS-SEC is 'managerial and professional'	Census	6.72
Inphhtype4	Logit of Proportion of households that are lone parent with all child(ren) non - dependent	Census	3.29
Inphhtype5	Logit of Proportion of households that are a couple with no children	Census	2.77
Inpltli	Logit of Proportion of people in households with a long-term limiting illness	Census	-3.40
ewpdccd12years	Proportion of people claiming Disability Living Allowance with a claim duration of 1-2 years	DWP	4.36
ewjsafemale	Proportion of females aged 16 and over claiming Job Seekers Allowance	DWP	-2.75
ewpdlaman	Proportion of people claiming Disability Living Allowance: Mobility Award Nil	DWP	2.35
Inispfemale	Logit of Proportion of females aged 16 and over claiming Income Support	DWP	-2.23
ewpcp25years	Proportion of people aged 60 and over	DWP	-1.99

	claiming Pension Credit, with a claim Duration of 2-5 Years		
Pflat	Percentage of household spaces that are a flat, maisonette or commercial building	Census	3.61
Inphhype7	Logit of Proportion of households that are a couple with all child(ren) non - dependent	Census	2.06
Inpltli_southeast	Interaction between Inpltli and southeast		-2.56
Inpltli_ewavhhpeop	Interaction between Inpltli and ewavhhpeop		2.42
ewpdccd12years_south	Interaction between ewpdccd12years and south		-2.34
ewpdccd12years_wales	Interaction between ewpdccd12years and wales		-2.47
ewpdlaman_Inphhype4	Interaction between ewpdlaman and Inphhype4		2.32

With no covariates included in the model the estimated residual area variance $\hat{\sigma}_u^2$ is 0.029(0.003) compared with 0.001(0.002) when the significant covariates are included in the model, a decrease of 96.06%. Therefore, these covariates together accounted for 96.06% of the total between area variance.

The most significant covariate in the model is the Census covariate 'phrpman', which has a T value of 6.72. As one would expect this covariate has a positive coefficient; as the proportion of the MSOA household reference persons aged 16 to 74 whose NS-SEC is managerial and professional increases so does the average weekly household income for that MSOA. 'ewpdccd12years' is the next most significant covariate in the model with a positive coefficient, and has a T-value of 4.36. It shows that as the proportion of people in an MSOA claiming Disability Living Allowance with a claim duration of 1-2 years increases so does the average weekly household income.

The most significant variable with a negative coefficient (-0.146) was 'Inpltli' which, has a T value of -3.4. This says that as the logit of the proportion of people living with long term limiting illness in an MSOA increases, the average weekly household income for that MSOA decreases.

The benchmarking ratios for this model are detailed in Table 5. These ratios are used to adjust the model-based estimates and their confidence intervals at the MSOA-level.

Table 5: Benchmarking results for net household weekly income (unequivalised)

Country/REGION	Aggregated model estimate	Survey estimate	Ratio of survey/model estimate
North East	491	529	0.93
North West	516	521	0.99
Yorkshire and the Humber	541	516	1.05
East Midlands	533	533	1.00
West Midlands	538	513	1.05
East of England	612	567	1.08
London	710	641	1.11
South East	757	629	1.20
South West	584	534	1.09
Wales	500	487	1.03

4.3.3 Net Weekly Household Income – Equivalised, Before Housing Costs

The model selected to estimate net weekly household income, equivalised, before housing costs was:

$$\begin{aligned}
 & 6.1039(0.01899) && \text{- Constant} \\
 & + 0.056(0.031) \text{northeast}_k \\
 & - 0.012(0.024) \text{northwest}_k \\
 & + 0.01(0.027) \text{york}_k \\
 & + 0.001(0.028) \text{eastmid}_k \\
 & - 0.029(0.025) \text{westmid}_k \\
 & - 0.015(0.026) \text{east}_k \\
 & + 0.008(0.022) \text{southeast}_k \\
 & - 0.022(0.026) \text{southwest}_k \\
 & - 0.043(0.030) \text{wales}_k \\
 & + 0.796(0.085) \text{phrpman}_k \\
 & + 0.741(0.204) \text{eghi}_l_k \\
 & + 0.167(0.025) \text{lnp16}_k \\
 & + 0.134(0.025) \text{lnphhtype4}_k \\
 & + 14.198(3.541) \text{ewpdccd12years}_k \\
 & - 0.034(0.014) \text{lnjsafemale}_k \\
 & + 0.171(0.053) \text{pflat}_k \\
 & - 0.978(0.150) \text{ewpcps}_k \\
 & - 2.163(0.815) \text{eghi}_l\text{_phrpman}_k \\
 & - 35.364(13.966) \text{ewpdccd12years_wales}_k
 \end{aligned}$$

} Regional/Country
 } Interactions

$$- 0.052(0.025) \text{ lnjsafemale_east}_k \quad \} \quad \text{Interactions}$$

$$+ u_j + e_{ij}$$

$$\hat{\sigma}_u^2 = 0.001(0.001)$$

$$\hat{\sigma}_e^2 = 0.310(0.004)$$

Equation [3]

Variables labelled as in Equation [1].

Table 6 contains a key to labels of the covariates. The covariates have been grouped by source.

Table 6: Key to covariates included in the model for net household weekly income, equivalised, before housing costs

Covariate Name	Label	Source	T ratio = $\left(\frac{\beta}{s.e} \right)$
Northeast	North East	Country/regional indicators	1.81
Northwest	North West	Country/regional indicators	-0.49
York	Yorkshire and The Humber	Country/regional indicators	0.38
Eastmid	East Midlands	Country/regional indicators	0.03
Westmid	West Midlands	Country/regional indicators	-1.17
East	East of England	Country/regional indicators	-0.58
Southeast	South East	Country/regional indicators	0.37
Southwest	South West	Country/regional indicators	-0.83
Wales	Wales	Country/regional indicators	-1.45
Phrpman	Proportion of HRPs aged 16 to 74 whose NS-SEC is 'managerial and professional'	Census	9.40
Eghi	Proportion/count of dwellings in English Council Tax bands G, H and I	VOA	3.64
lnp16_59	Logit of Proportion of people aged 16 to 59		6.70

Inphhtype4	Logit of Proportion of households that are lone parent with all child(ren) non - dependent	Census	5.42
ewpdccd12years	Proportion of people claiming Disability Living Allowance with a claim duration of 1-2 years	DWP	4.01
Lnjsafemale	Logit of Proportion of females aged 16 and over claiming Job Seekers Allowance	DWP	-2.38
Pflat	Percentage of household spaces that are a flat, maisonette or commercial building	Census	3.24
Ewpcps	Proportion of single people aged 60 and over claiming Pension Credit	DWP	-6.50
eghi_phrpman	Interaction between eghi and phrpman		-2.65
ewpdccd12years_wales	Interaction between ewpdccd12years and wales		-2.53
Lnjsafemale_east	Interaction between Lnjsafemale and east		-2.02

With no covariates included in the model the estimated residual area variance $\hat{\sigma}_u^2$ was 0.026(0.002) compared with 0.001(0.001) when the significant covariates were included in the model, a decrease of 95.54%. Therefore, these covariates together accounted for 95.54% of the total between area variance.

The most significant covariate in the model is the Census covariate, 'phrpman', which has a T value of 9.4. This covariate represents the proportion of people in a MSOA who are in a managerial or professional occupation. Higher levels of this occupation type relate to higher incomes. Another highly significant covariate is 'lnp16_59', with a T value of 6.7. This shows that as the proportion of the MSOA population who are aged 16-59 increases, average weekly household income for that MSOA.

The benchmarking ratios for this model are detailed in Table 7. The ratios in Table 7 are used to adjust the model-based estimates and their confidence intervals at the MSOA-level.

Table 7: Benchmarking results for net household weekly income before housing costs, equivalised

Country/REGION	Aggregated model estimate	Survey estimate	Ratio of survey/model estimate
North East	470	502	0.94
North West	492	494	1.00
Yorkshire and the Humber	505	487	1.04
East Midlands	502	499	1.01
West Midlands	490	481	1.02
East of England	569	539	1.05
London	658	591	1.11
South East	681	580	1.17
South West	551	515	1.07
Wales	469	473	0.99

4.3.4 Net Weekly Household Income – Equivalised, After Housing Costs

The model selected to estimate net weekly household income, equivalised, after housing costs was:

$$\begin{aligned}
 & 5.919(0.021) && \text{- Constant} \\
 & + 0.050(0.034) \text{northeast}_k \\
 & - 0.023(0.026) \text{northwst}_k \\
 & + 0.017(0.030) \text{york}_k \\
 & + 0.002(0.031) \text{eastmid}_k \\
 & - 0.030(0.029) \text{westmid}_k \\
 & - 0.048(0.030) \text{east}_k \\
 & - 0.012(0.026) \text{southeast}_k \\
 & - 0.013(0.033) \text{southwst}_k \\
 & - 0.044(0.033) \text{wales}_k \\
 & - 0.859(0.141) \text{ewpcptotal}_k \\
 & + 0.926(0.101) \text{phrpman}_k \\
 & + 0.137(0.030) \text{lnphhtype4}_k \\
 & + 0.099(0.038) \text{lnp16}_k \\
 & - 0.035(0.018) \text{lnjsafemale}_k \\
 & + 12.310(4.242) \text{ewpdccd12years}_k \\
 & - 0.054(0.018) \text{lnpftstud}_k \\
 & - 0.071(0.021) \text{lnisptotal}_k \\
 & + 0.067(0.024) \text{lnpdccd25years}_k \\
 & - 0.099(0.031) \text{lnjsafemale}_k \\
 & - 0.080(0.032) \text{lnisptotal}_k \\
 & + 0.109(0.047) \text{lnpftstud}_k \\
 & - 0.330(0.152) \text{phrpman}_k \\
 & + u_j + e_{ij}
 \end{aligned}$$

} Regional/Country
} Interactions

$$\hat{\sigma}_u^2 = 0.003(0.002)$$

$$\hat{\sigma}_e^2 = 0.433(0.005)$$

Equation [4]

Variables labelled as in Equation [1].

Table 8 contains a key to the labels of the covariates.

Table 8: Key to covariates included in the model for equivalised net household weekly income after housing costs

Covariate Name	Label	Source	T ratio = $\left(\frac{\beta}{s.e}\right)$
northeast	North East	Country/regional indicators	1.46
northwst	North West	Country/regional indicators	-0.88
york	Yorkshire and The Humber	Country/regional indicators	0.58
eastmid	East Midlands	Country/regional indicators	0.07
westmid	West Midlands	Country/regional indicators	-1.04
east	East of England	Country/regional indicators	-1.59
southeast	South East	Country/regional indicators	-0.47
southwst	South West	Country/regional indicators	-0.39
wales	Wales	Country/regional indicators	-1.36
ewpcptotal	Proportion of people aged 60 and over claiming Pension Credit	DWP	-6.11
phrpman	Proportion of HRPs aged 16 to 74 whose NS-SEC is 'managerial and professional'	Census	9.18
lnphhtype4	Logit of Proportion of households that are lone parent with all child(ren) non -dependent	Census	4.60
lnp16_59	Logit of Proportion of people aged 16 to 59		2.62
lnjsafemale	Logit of Proportion of females aged 16 and over claiming Job Seekers Allowance	DWP	-1.92
ewpdccd12years	Proportion of people claiming Disability Living Allowance with a claim duration of 1-2 years	DWP	2.90
lnpftstud	Logit of Proportion of people aged 16 to 74 who are full-time students	Census	-2.91
lnisptotal	Proportion of people aged 16 and over claiming Income Support	DWP	-3.39
lnpdccd25years	Logit of Proportion of people claiming Disability Living Allowance with a claim duration of 2-5 years	DWP	2.77
lnjsafemale_east	Interaction between lnjsafemale and east		-3.19
lnisptotal_lnp16_59	Interaction between lnisptotal and lnp16_59		-2.51
lnpftstud_southwst	Interaction between lnpftstud and southwst		2.33
phrpman_westmid	Interaction between phrpman and westmid		-2.17

With no covariates included in the model the estimated residual area variance $\hat{\sigma}_u^2$ was 0.037(0.003) compared with 0.003(0.002) when the significant covariates were included in the model, a decrease of 91.12%. Therefore, these covariates together accounted for 91.12% of the total between area variance.

The most significant covariate in the model is the census covariate ‘phrpman’ that has a T value of 9.18. As the proportion of the MSOA population whose occupation is managerial or professional increases so does the average weekly household income for that MSOA. The covariate ‘lnphhtype4’ has a T value of 4.6 which shows that as the proportion lone parent households in the MSOA that are with all children non -dependent increases, average weekly household income for that MSOA also increases.

The benchmarking ratios for this model are detailed in Table 9. The ratios in Table 9 are used to adjust the model-based estimates and their confidence intervals at the MSOA-level.

Table 9: Benchmarking results for net household weekly income, equivalised, after housing costs

Country/REGION	Aggregated model estimate	Survey estimate	Ratio of survey/model estimate
North East	420	459	0.91
North West	439	452	0.97
Yorkshire and the Humber	456	456	1.00
East Midlands	454	470	0.96
West Midlands	438	444	0.99
East of England	502	496	1.01
London	538	480	1.12
South East	607	529	1.15
South West	486	473	1.03
Wales	426	435	0.98

4.3.5 Observations

As one would expect the four models are very similar. Although some of the covariates may be different between the four equations the models are generally explaining the same MSOA characteristics. The models include covariates from the following list.

1. In all four models the most significant covariate is a Census covariate indicating socio-economic class (phrpman). As one would expect this covariate has a positive coefficient, meaning as the proportion of adults or household representatives whose occupation is classified as managerial or professional increases so does the average weekly household income for that MSOA.
2. The majority of regional and country indicators in each model are not significant but are included since benchmarking is carried out at this level.
3. The final types of covariates included in the models are interaction effects. The majority of interaction terms involve regional and country indicators. This shows that some covariates have different effects in different regions and for Wales.

Some of these results described may be unexpected, however, it should be remembered that the relationships observed should not be taken in isolation but alongside the other relationships described by the other covariates present in the model.

The four models show different benchmarking ratios but this is not a surprising result since all region/country terms are included in each model. Investigations have shown that any difference between the aggregated model-based estimate and the direct survey estimate are caused either by using the transformation and modelling on the logit scale or because survey weights are used in the calculation of the survey estimates but not considered in the modelling process (see Appendix A for more details). The pattern in the ratios is generally the same between the four income types; the greatest discrepancies between the aggregated model based estimates and the survey estimates occur in the London and the South East Regions.

5. Results of Modelling for Income

5.1 Total Weekly Household Income (unequivalised)

The estimates of total weekly household income (unequivalised) for 2013/14 were produced using the right hand side of Equation [1] in Section 4.3.1 (excluding the u_j and e_{ij} terms) by substituting the known values of the MSOA covariates in to the fitted model.

Figure 2 provides a visualisation of the model-based estimates and their 95% confidence intervals. Unlike confidence intervals for direct estimates that can be interpreted in terms of repeated sampling; the confidence intervals for model-based estimates represent the uncertainty in the modelling process. This means that about 95% of MSOAs have a true value contained in the confidence interval. 95% confidence intervals given by ± 1.96 standard error are used where the standard error includes between postcode sector variance and parameter estimate variance. The equations for the confidence intervals are detailed in Appendix A and these confidence intervals tell us that we can be 95% certain that the true value of average household income lies between the upper and lower confidence limits.

Figure 2: Model MSOA estimates and 95% Confidence Intervals for total weekly household income (unequivalised)

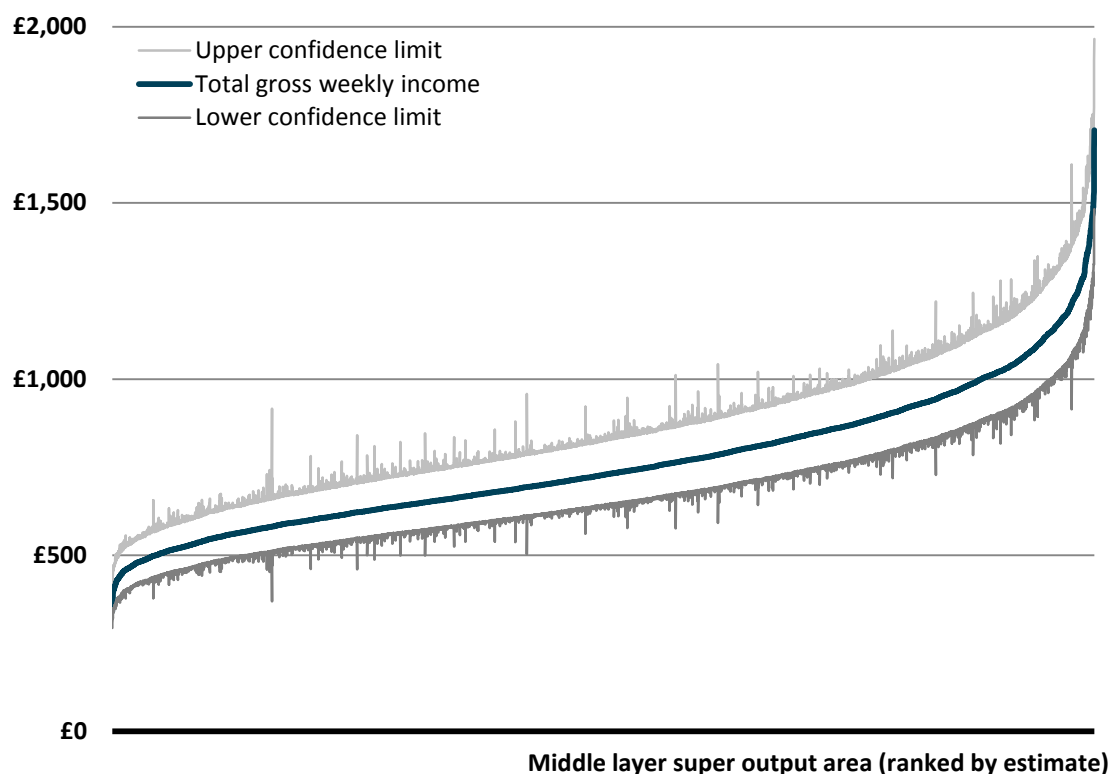
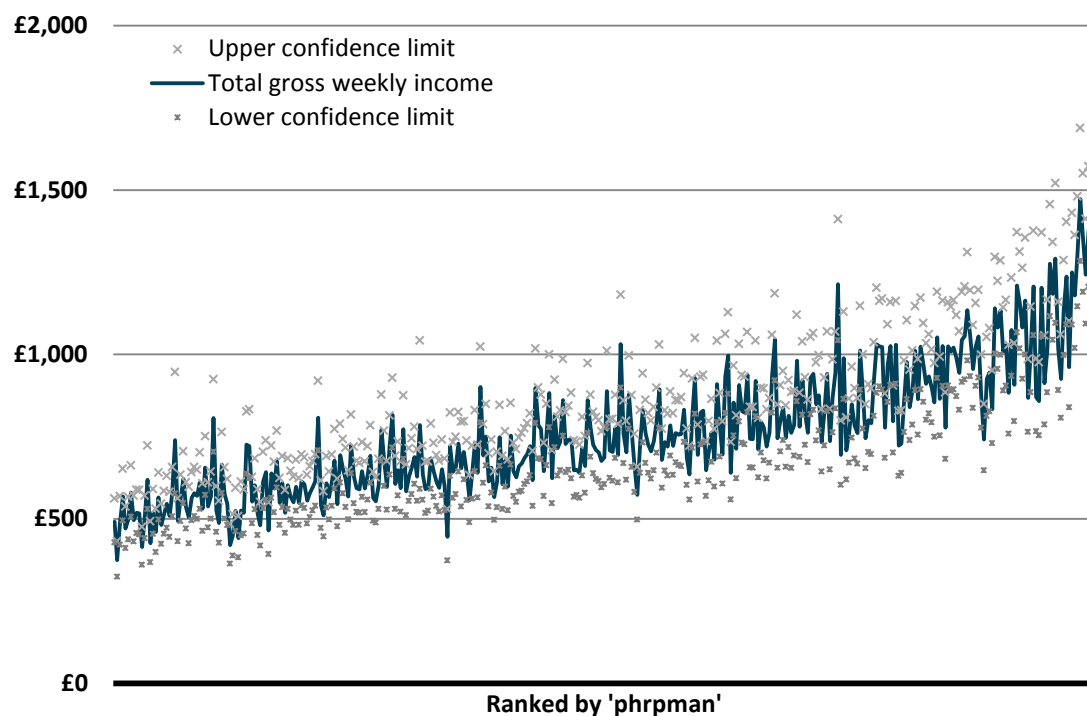


Figure 3: Sample of model MSOA estimates and 95% Confidence Intervals for total income (unequivalised)



In order to aid clear illustration Figure 3 displays a random sample (about 5%) of all the MSOA estimates and their confidence intervals. The estimates are ranked by the value of the Census covariate 'phrman', the proportion of people aged 16 to 74 in a MSOA whose NS-SEC is 'managerial and professional'. The chart shows how this covariate has a strong relationship with income and hence explains a lot of the variance. As the value of the covariate increases so does the value of the model-based estimate.

Map 1 shows the coefficient of variation (CV) for total weekly income and map 2 shows the asymmetry of the confidence intervals for total weekly income by MSOA in 2013/14. A large majority of MSOAs (5,650) have CVs between 6.17% and 7.24%. This group was defined by calculating natural Jenks breaks, commonly used in choropleth mapping. The method derives 'natural' breaks in the data by identifying clusters of observations which include a large number of data points relative to the values immediately above and below each cluster, or class. Map 1 shows that there were a relatively small number of MSOAs in the class containing the largest CVs. This class contains 34 MSOAs which had CVs of between 10.24 and 24.04. These MSOAs do not appear to be clustered in any particular geographical area or region.

The model used to estimate income, and its confidence intervals, uses the exponential function, having been back-transformed from the natural log scale. As such, the confidence intervals are asymmetrical and the upper confidence limits tend to lie further from the estimate than the lower confidence limits. A large majority of MSOAs (6,783) had upper confidence limits that lie between 1.5% and 2.1% of the estimate further away from the estimate than the lower confidence limit. This class of MSOAs is distributed across England and Wales. A small class of MSOAs (26) had relatively large asymmetry in their confidence intervals, ranging from 4.7% to 21.1% of their estimates of income. These MSOAs do not

appear to be clustered in any particular geographical area or region.

Map 1 (left): Coefficient of variation - Total weekly income by MSOA, England and Wales, 2013/14;

Map 2 (right): Distance further from estimate of total weekly income of the upper confidence limit than the lower confidence limit (expressed as a percentage of the estimate) by MSOA, England and Wales, 2013/14



Source: Office for National Statistics and Ordnance Survey under the Open Government Licence v3.0. Contains OS data © Crown copyright 2016

Similarly small numbers of MSOAs are in the highest classes of both CVs and asymmetry for the three other modelled estimates. These can be seen in Map 3 to Map 8 in Appendices D and E along with the confidence interval plots for the other variables.

5.2 Summary of Results

The figures for the model-based estimates and their confidence intervals show broadly similar widths in confidence intervals for the four types of income estimates. Commentary describing the distribution and geographical features of the four income estimates variables is available in the statistical bulletin accompanying these statistics. MSOAs with higher income estimates have wider confidence intervals than lower ones in absolute terms, although they are not proportionately larger.

A different selection of covariates (appropriate for each income type) has caused some differences in the resulting estimates. Slight inconsistencies (when examining point estimates) may occur between the income types for particular MSOAs but the models selected are the best possible to model the general patterns of income over all MSOAs. This reinforces the need to look at the confidence interval for the income estimates not just the point estimate since the confidence intervals summarise the variability in the estimates caused by the modelling process, see Chapter 8.

6. Quality of the Estimates

Introduction

This chapter describes the different diagnostic checks that have been used to assess the appropriateness of the models developed. The diagnostic checks employed here are those developed by ONS for small area estimation (Brown et al (2001)) as well as some additional ones. Each diagnostic test is described and the results displayed for modelling total weekly household income (unequalised) for England and Wales. The results for England & Wales for all four income types are summarised at the end of the chapter. All tables and plots are displayed in Appendix E. In this chapter we describe the results for **total income only**.

6.1 Residual vs. Model Estimates Diagnostic Plot

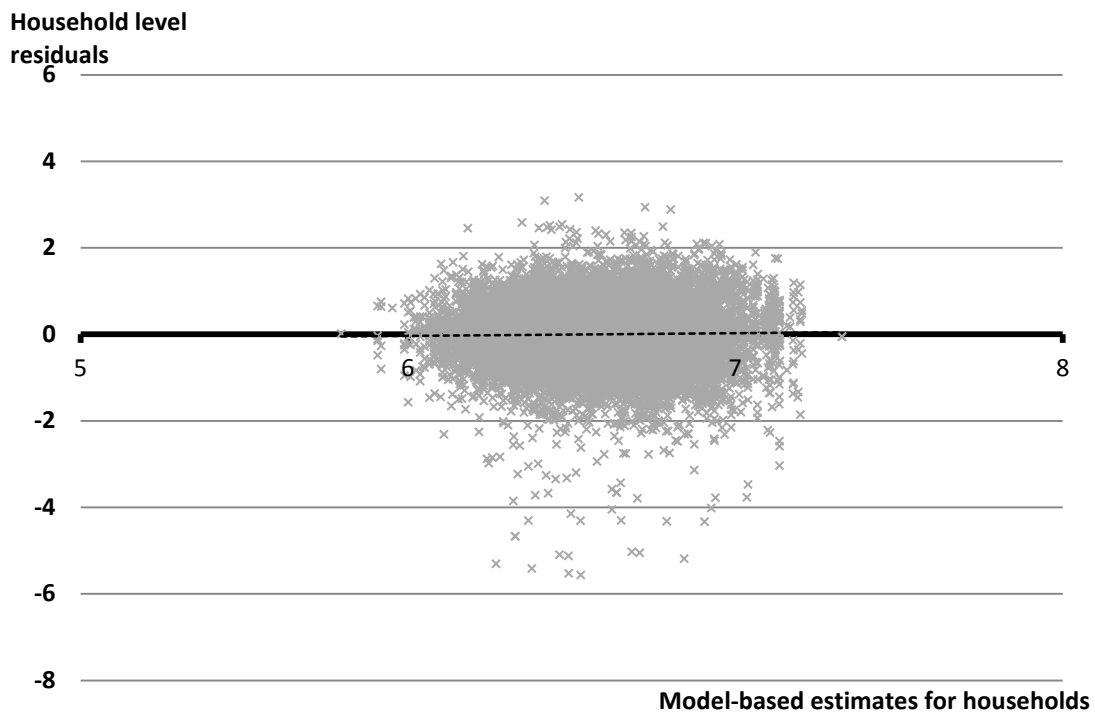
A plot of model estimates against model residuals both at the household and the area level is a method of checking that the model assumptions are satisfied and the model accurately describes the population. Here we are testing for two things: model mis-specification and non-constant variance of the residuals (heteroscedasticity). If any pattern remains in the residuals this implies model mis-specification e.g. a covariate influential to income has been left out of the model. We require constant variance in the area level residuals since this will have an impact on the calculation of the confidence intervals.

Model estimates are calculated at the household level (on the natural log (ln) scale) and plotted against the household level residuals, e_{ij} 's in. The equation of the regression line is shown below. The standard errors (displayed in parentheses) can be used to determine whether the constant and linear terms are significantly different from 0. This equation shows that the regression line is significantly different to the line $y = 0$:

$$y = -0.443(0.181) + 0.067(0.028)x$$

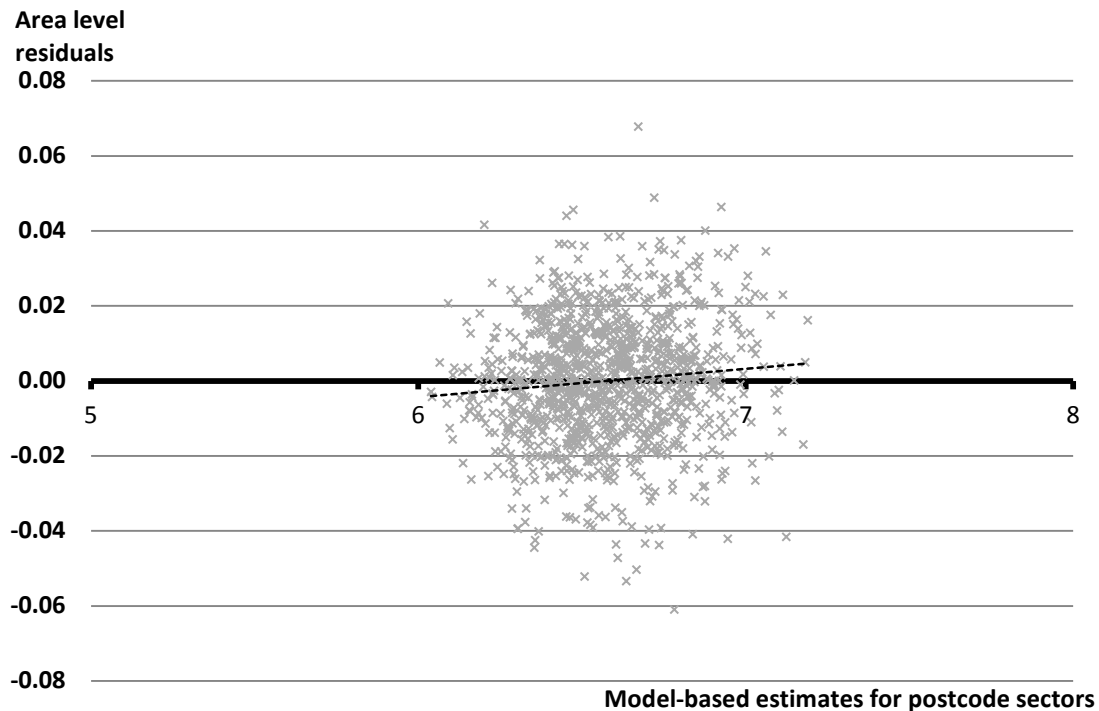
However, the line does not deviate sufficiently from $y = 0$ to cause concern over the quality of the model. This is shown in Figure 10 in which there is no obvious pattern to the residuals.

Figure 10: Household level residuals against model estimates, total income (unequivalised)



Due to the structure of the model area level residuals, $u_j^i S$, refer to postcode sectors (PCS). For the plot of area level residuals we require model-based estimates at the PCS level, however, covariates are by MSOA and not PCS. In order to form model-based estimates for PCS for the plot an approximate method is used. A weighted average (using the survey weights) of the household model-based estimates is calculated at the PCS level. For this residual diagnostic we are making the assumption that the results at PCS level would highlight any problems at the MSOA-level. Note that these PCS model estimates are not calibrated and are displayed on the ln scale. Figure 11 displays the area level residual plot.

Figure 11: Area level residuals against model-based estimates, total income (unequivalised)



The equation of the regression line shown below is significantly different to the line $y=0$:

$$y = -0.050(0.015) + 0.008(0.002)x$$

For gross income, the plot shows evidence of a marginal pattern in variance as estimates increase. The constant variance assumption does not strictly hold after modelling both at the household and the area level, however, the results from the other diagnostics support the model.

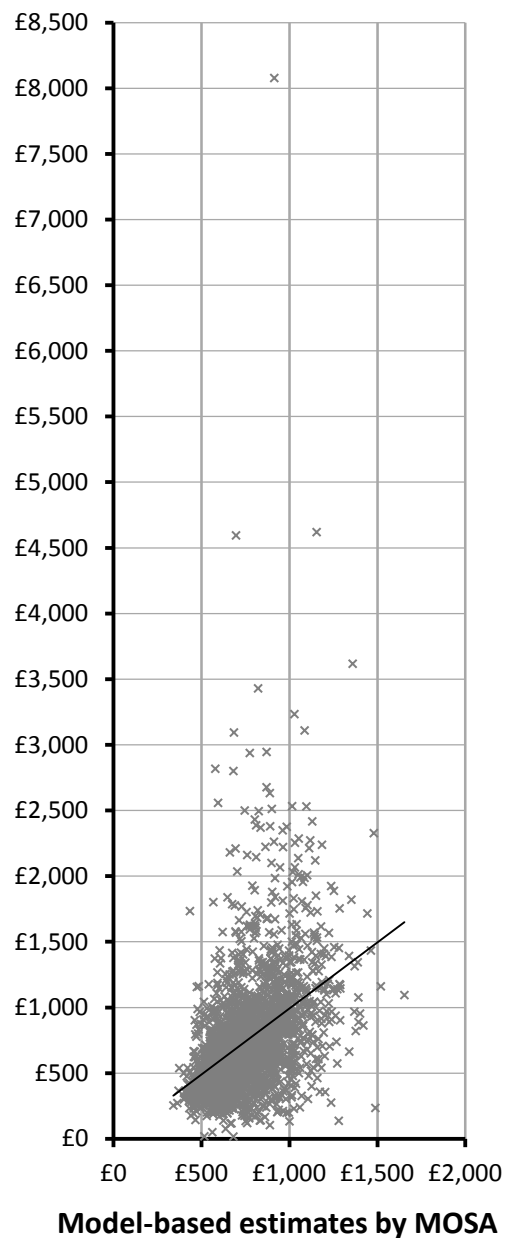
6.2 Model vs. Sample Estimates Diagnostic Plot

A plot of direct survey estimates (y-axis) against model-based estimates (x-axis) for MSOAs for which there is a sample, is one method of assessing whether the relationship between the target variable and the covariates has been specified properly. For good model-based estimates, the direct estimates will be randomly distributed around the estimates and the regression line between the two will be very close to the line $y=x$. If the relationship between the target variable and the covariates has been mis-specified or mis-estimated then the relationship between the direct and model-based estimates would be expected to be curved or possibly scattered round a different straight line than the $y=x$ line. An important assumption when using this diagnostic is that the direct estimates are unbiased. The technique for calculating direct survey estimates at a MSOA-level is described in Appendix F.

Figure 12 shows the plot of direct survey estimates against model MSOA estimates for total weekly household income (unequivalised). The linear regression line (solid line) is shown compared with the $y=x$ line (dashed line).

Figure 12: Model-based estimates vs. sample estimates, total weekly household income (unequivalised)

Direct survey estimates



The plot shows a much wider variation in direct survey estimates than for the model-based estimates. This is due to the fact that the points represent data for MSOAs, which will have an extremely small sample size.

The equation of the linear regression line is (the standard errors are displayed in parentheses):

$$y = -15.717(33.639) + 1.007(0.043)x.$$

A quadratic curve was also fitted to the points:

$$y = -163.553(110.529) + 1.386(0.273)x - 0.0002(0.0002)x^2.$$

The result shows that in quadratic fit, the quadratic term is not significant and neither is the intercept. In the linear fit, the intercept term is not significant and the slope term is not significantly different from one. Thus the fit is very close to the $y=x$ line. This shows that at least in sampled areas the modelled estimates do not show signs of bias.

6.3 Coverage Diagnostic

The purpose of this diagnostic is to examine the validity of the confidence intervals for the model-based estimates. For those MSOAs in sample, there will be direct survey estimates with associated 95% confidence intervals (described in Appendix F). The diagnostic measures the overlap between the direct confidence intervals and the corresponding model-based estimate confidence intervals, i.e. it measures the percentage of MSOAs for which the model and direct confidence intervals overlap.

However, the overlap between two independent 95% confidence intervals for the same quantity is higher than 95%, therefore it is necessary to modify the nominal coverage levels (i.e. narrow the width) of the confidence intervals being compared to ensure a 95% overlap.

The modification is based on the fact that if X and Y are two independent normal random variables, with the same mean but with different standard deviations, σ_X and σ_Y respectively then the standard deviation of the difference is $\sqrt{\sigma_X^2 + \sigma_Y^2}$. If $z(\alpha)$ is such that the probability that a standard normal variable takes values greater than $z(\alpha)$ is $\alpha/2$, (eg $\alpha=0.05$ and $z(\alpha)=1.96$ for a 95% confidence interval under a normal distribution) then a sufficient condition for there to be probability of α that the two intervals $X \pm z(\beta)\sigma_X$ and $Y \pm z(\beta)\sigma_Y$ do not overlap is when

$$\begin{aligned} z(\beta) &= z(\alpha) \frac{\sqrt{\sigma_Y^2 + \sigma_X^2}}{\sigma_Y + \sigma_X} \\ &= z(\alpha) \left(1 + \frac{\sigma_X}{\sigma_Y}\right)^{-1} \sqrt{1 + \frac{\sigma_X^2}{\sigma_Y^2}}. \end{aligned}$$

Consequently, this diagnostic takes $z(\alpha) = 1.96$, calculates $z(\beta)$ using the above formula, with σ_X replaced by the estimated standard error of the model-based estimate and σ_Y replaced by the estimated standard error of the direct estimate and then computes the

overlap proportion between the corresponding $z(\beta)$ -based confidence intervals. For $z(\alpha) = 1.96$ this proportion should be 95%. Any significant deviation from a 95% overlap will indicate that the model based confidence intervals are generally too wide or too narrow.

The analysis shows that an overlap occurs in all of the 2,539 MSOAs (which is greater than the required 95%). A pooled variance has been used to calculate the confidence intervals for the direct estimates (see Appendix F) and this will result in an overestimation of these confidence intervals and hence a coverage percentage slightly greater than 95% is not a surprising result.

6.4 Wald Statistic

This diagnostic test assesses the assumptions underlying the model by using a Wald goodness of fit statistic to test whether there is a significant difference between the expected values of the direct estimates and the model-based estimates. Typically, small area-level model-based and direct survey estimates will be approximately correlated. Consequently, a Wald statistic for testing the MSOA-level goodness-of-fit of a model-based set of estimates is:

$$W = \sum_j \frac{(z_j - \zeta_j)^2}{V(z_j) + V(\zeta_j)}.$$

where ζ_j is the model-based estimate of the average weekly household income for MSOA j , $V(\zeta_j)$ is its estimated variance and z_j and $V(z_j)$ are the corresponding direct MSOA estimate and variance. We assume the covariance $C(z_j, \zeta_j)$ is negligible. Under the hypothesis that the model-based estimates are equal to the expected values of the direct estimates, and provided the sample sizes in the MSOAs are sufficient to justify central limit assumptions, W will then have a χ^2 distribution with degrees of freedom equal to the number of MSOAs in the population.

The goodness-of-fit statistic for the model developed here is 561.1 on 2,539 degrees of freedom, this has a p-value of 1.00. There is no significant evidence to reject a χ^2 distribution. Therefore, there is no significant difference between the expected values of the model-based estimates and the direct survey estimates.

6.5 Stability Analysis

This diagnostic test analyses the stability of the model's predictive power. The data are split at random to obtain two data sets; Data_A and Data_B . The data are split in such a way to ensure as much as possible that the two data sets are the same in terms of size and MSOAs represented. The model is fitted to one half of the data, Data_A , to obtain the regression coefficients $\hat{\beta}_{k_A}$. In a similar way Data_B is used in the model to obtain the regression coefficients $\hat{\beta}_{k_B}$. These two sets of regression coefficients are then used to obtain two sets of comparable model-based estimates for all MSOAs. This process is repeated 10 times and

for each repetition the difference between the two sets of estimates is measured to evaluate the stability of the model.

A relative root mean square error (RRMSE) as defined below is also used as a measure of how close the two sets of model-based estimates are. A small RRMSE indicates that the differences between the two sets of estimates are not significant.

$$RRMSE = \sqrt{\left(\sum_i \frac{1}{n} \left(\frac{\hat{Y}_B - \hat{Y}_A}{\hat{Y}_A} \right)^2 \right)}$$

where \hat{Y}_A and \hat{Y}_B are the model-based estimates calculated using regression coefficients $\hat{\beta}_{k_A}$ and $\hat{\beta}_{k_B}$ respectively and n is the total number of MSOAs. For total weekly household income (unequalised) the median RRMSE for the 10 repetitions is 0.047.

The RRMSE shows that the two sets of estimates are fairly similar and that there is stability in the model. An RRMSE of greater than 0.5 is considered here as an indication of instability.

Figure 13: Model-based estimates from stability analysis, total weekly household income

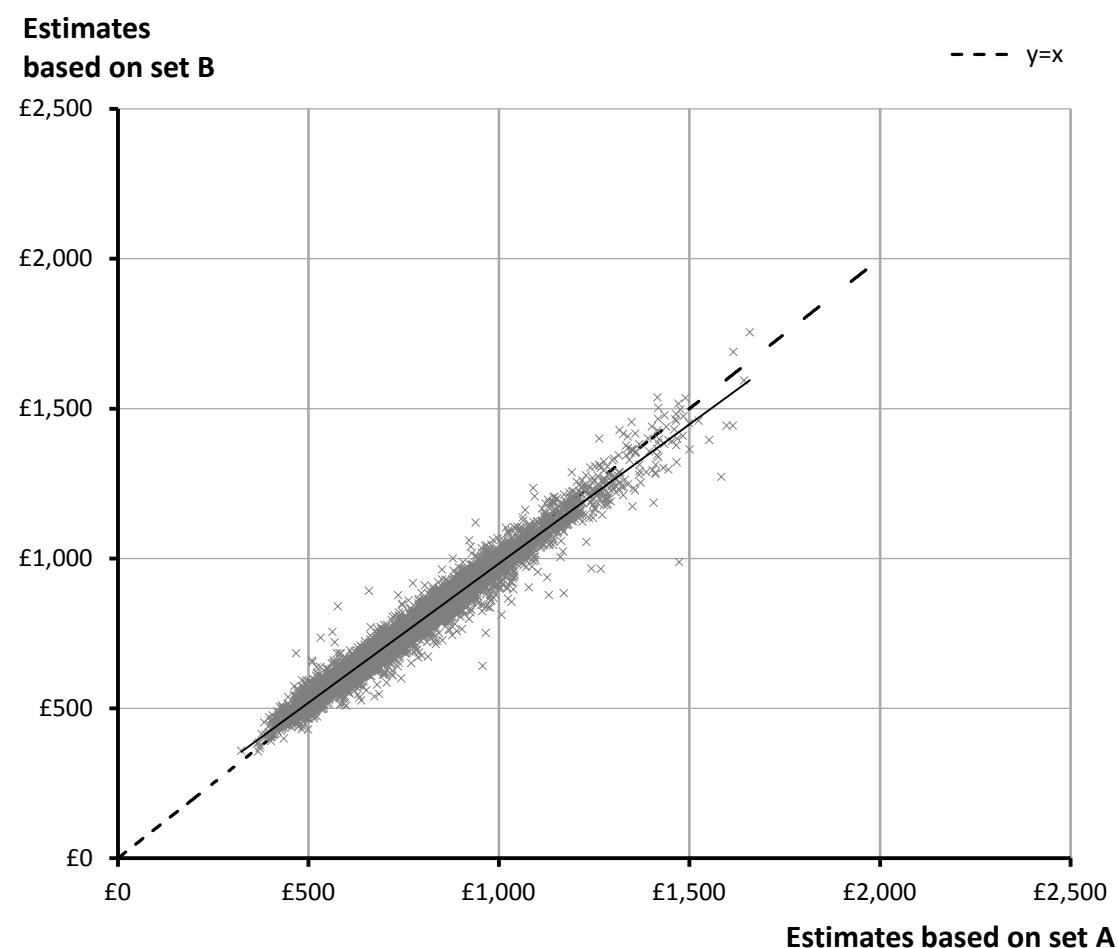


Figure 13 shows a comparison of the model-based estimates obtained with the $y=x$ line, for one set of estimates.

6.6 Diagnostic Results

Table 10 summarises the results of the standard diagnostic tests for all four income types for England and Wales. All of the tables and plots associated with the results can be found in Appendix E.

Table 10: Diagnostic results for all four income types estimated, England and Wales

Diagnostic Measure		Total Weekly Household Income (unequalised)	Net Weekly Household Income (unequalised)	Net Weekly Household Income, Equalised Before Housing costs	Net Weekly Household Income, Equalised After Housing costs
Residual vs Model Estimates Household Level Residuals	Constant (SE) Slope (SE)	-0.443(0.181) 0.067(0.028)	-0.211(0.190) 0.033(0.030)	-0.250(0.173) 0.040(0.028)	-0.473(0.170) 0.077(0.028)
Residual vs Model Estimates Area Level Residuals	Constant (SE) Slope (SE)	-0.050(0.015) 0.008(0.002)	-0.012(0.007) 0.002(0.001)	-0.024(0.008) 0.004(0.001)	-0.072(0.016) 0.012(0.003)
Model vs Sample Estimates	Constant (SE) Slope (SE)	-15.717(33.639) 1.007(0.043)	58.250(23.032) 0.844(0.037)	16.577(21.714) 0.927(0.038)	41.891(20.600) 0.882(0.041)
Model vs Sample Estimates	Constant (SE) Slope (SE) Quadratic term (SE)	-163.553(110.529) 1.385(0.273) -0.0002(0.0002)	- 115.521(78.548) 1.401(0.244) -0.0004(0.0002)	184.681(84.804) 0.339(0.289) 0.0005(0.0002)	7.752(69.284) 1.019(0.269) -0.0001(0.0003)
Coverage	%	100	99.92	100	100
Wald	P-value	1.000	1.000	1.000	1.000
Stability Analysis	RRMSE	0.047	0.052	0.044	0.056

6.7 Conclusions

Some of the plots of the household and area level residuals for each income type (Appendix E) indicate that a slight pattern remains in the data after modelling, and for these the modelling assumptions are not fully satisfied. The area level residual plots show that the area level residuals do not have constant variance for three of the models. However, where both of the residual plots did show a pattern, the bias plots show that the regressions between the direct and model-based estimates are close to $y = x$, and the coefficient of a quadratic term is not significantly different from zero.

The coverage diagnostic shows coverage greater than 95% in all four models indicating that the confidence intervals of the model-based estimates are possibly conservative. (This means that the true value of mean income would be within the confidence interval for more than 95% of the MSOAs). However this may also be due to over estimating the variances for the direct estimates. For all four models the Wald goodness-of-fit statistic shows no significant difference between the expected value of the direct and model-based estimates. The stability analyses for the four models indicate that the different sets of data produce similar sets of estimates.

In conclusion the analysis in this chapter shows that in general the models for England and Wales are well specified and assumptions are satisfied. This provides confidence in the accuracy of the estimates and their confidence intervals produced from the models.

7. Comparing results for 2011/12 and 2013/14, and measuring change

MSOA-level model-based estimates of average weekly household income have been produced for 2013/14 in England and Wales, fulfilling users' requirements for income information at MSOA level.

7.1 Models

In 2011/12 each model related the FRS survey estimate of weekly household income to the following covariates:

- Average number of people per household
- Proportion of people aged 16-74 whose NS-SEC is 'managerial and professional'
- Proportion of people claiming Disability Living Allowance: Mobility Award Higher
- Transactions by Dwelling Type; Total Sales
- Proportion of people aged 60 and over claiming Pension Credit
- Proportion of people in households with a long-term limiting illness
- Region/country in which MSOA lies

In 2013/14 each model contained the following covariates:

- Proportion of people claiming Disability Living Allowance with a claim duration of 1-2 years
- Proportion of females aged 16 and over claiming Job Seekers Allowance
- Proportion of households that are lone parent with all child(ren) non-dependent
- Region/country in which MSOA lies

7.2 Diagnostics

Some plots of household and area level residuals for all models for both 2011/12 and 2013/14 showed a slight pattern in the data after modelling. However, where there were patterns with the residual plots, the

plots of the modelled estimates against the direct estimates showed little or no bias. The bias plots for 2013/14 were not as close to the $y=x$ line as the bias plots for 2011/12.

For both years, the coverage diagnostic shows coverage greater than 95% for all four models indicating that the confidence intervals of the model-based estimates are possibly conservative. However, this may be due to over estimating the variances for the direct estimates. For both time periods and all models the Wald goodness-of-fit statistic shows no significant difference between the expected value of the direct and model-based estimates. Also, the stability analyses for both time periods indicate that the different sets of data produce similar sets of estimates for all four of the models.

The diagnostics for the 2011/12 and 2013/14 models produce fairly consistent results indicating that in general the models for England and Wales are well specified and the assumptions are satisfied. This produces confidence in the accuracy of the estimates and their confidence intervals produced from the models.

7.3 Estimates

The different models described above have been independently chosen to give the best point-in-time estimates of household income for the appropriate time period and for the appropriate geography. In particular the synthetic estimation methodology, by borrowing strength nationally, tends to draw estimates at the low and high ends of the distribution towards the national mean. This is an acceptable drawback for point-in-time estimation as it is more than compensated by the advantages of borrowing strength nationally in increasing estimate precision. However it is problematic when the focus is on measuring local area change. This is because the small area estimate of change is drawn towards the national mean of change and no longer picks out local variability which in many cases is what is of particular interest. For this reason the synthetic estimation applied here is not optimised to give the best estimate of local change.

7.3.1 Covariates

The following covariates were available for modelling the 2013/14 and 2011/12 MSOA model-based estimates of income.

Table 11: Covariate data used in the models for the 2011/12 and 2013/14 model-based estimates of income

2011/12	2013/14
Census data, 2011	Census data, 2011
HMRC data, 2011	HMRC data, 2013
DWP benefit data, 2011/12	DWP benefit data, 2013
Region/country indicators	Region/country indicators
CLG dwelling prices (for changes of ownership), 2009(Jan-Dec)	ONS House Price Statistics for Small Areas year ending March 2014
Council Tax data, March 2011	Council Tax data, March 2013
	DECC Energy Consumption data 2013

Table 11 shows that different covariate data sets were available and used at the time of modelling the 2011/12 and 2013/14 model-based estimates of average income.

Different covariates have been selected in the models for 2011/12 and 2013/14. This is both a consequence of the covariate selection process as well as the availability of different covariate data sets for the two time periods. The covariate selection procedure ensures that only covariates strongly related to income are selected for each model. However, as a consequence of the selection of different covariates, sharp changes in the estimates for particular areas could result. A difference in the estimates for an MSOA between 2011/12 and 2013/14 could partly be due to differences in the covariates selected in the models rather than a real change in the mean household income for that area.

7.3.2 Geography of estimation

The 2001/02 income estimates were produced on 2003 CAS wards but more recent estimates are produced on MSOAs. Only 924 of 8,850 CAS wards are directly equivalent to MSOAs (of which there are 7,201), i.e. the majority of CAS wards are physically different to MSOAs. Comparisons between 2001/2 estimates and later estimates are therefore not usually possible because of boundary differences.

7.4 Estimates of change

To enable comparisons between two sets of model-based estimates the methodology employed should be the same, as should the output geographies for the estimates. The method used to produce the 2004/05, 2007/8, 2011/12 and 2013/14 model-based estimates is the same and all sets of estimates refer to MSOA boundaries. Therefore, it is possible to draw comparisons between estimates for the same MSOA in the two different time periods. However, the 2011/12 and 2013/14 estimates use the 2011 Census geography which contains more MSOAs and some altered MSOA boundaries than previous estimates, which were based on the 2001 Census geography. Therefore, for some MSOAs direct comparisons of income between the 2013/14 or 2011/12 estimates and earlier estimates is not possible. In these instances, the geography code which represents the MSOA for 2013/14 (or 2011/12) will not match with any geography code for earlier estimates.

If the confidence intervals for the estimates at different time periods do not overlap then there is some evidence of change over time but, users are warned not to interpret the difference between the point estimates as a precise measure of change.

Each estimate has been independently produced as the best estimate of mean household income at the appropriate point in time but as such they are not optimised to give the best measure of change. The selection of different covariates for previous models may induce changes in the estimates for particular areas where no underlying change has actually taken place.

ONS is aware that there is a strong user interest in development of a more efficient measure of change.

8. Guidance on the Use of the Estimates

The results of the diagnostic checks presented in Chapter 6 show that the models are well specified and the modelling assumptions generally hold. However, in the use of the model-based estimates, one needs to be aware of possible limitations. The quality of the estimates is strongly dependent upon the quality and relevance of the input data sources (covariates) used and the fit of the model achieved. In this particular case, the estimates are produced using the most up-to-date covariate data sources to match the 2013/14 survey data. Hence the estimates should be fully consistent with the current profile of the area.

In common with any ranking based on estimates, when ranking MSOAs by income, care must be exercised in interpreting the ranking of the MSOAs. One needs to take into account the variability of the estimates when using these figures. For example, the confidence interval around the highest ranked MSOA suggests that the estimate lies among the group of MSOAs with the highest income levels rather than being the MSOA with the highest average MSOA income. Estimates for two particular MSOAs can be described as significantly different if the confidence intervals for the estimates do not overlap.

Although these model-based estimates can be used to rank MSOAs by income they cannot be used to make any conclusions on the distribution of income over the MSOAs. The estimation procedure will tend to shrink estimates towards the average level of income for the whole population so estimates at each end of the

scales tend to be over or under estimated.

Nevertheless estimates can be used to make inferences such as the average weekly household income for MSOA A is greater than the value for MSOA B (if the appropriate confidence intervals do not overlap).

The model-based methodology produces MSOA-level estimates of average weekly income. These MSOA-level estimates can be aggregated to provide income estimates for larger geographical areas such as Local Authority Districts (LADs) or regions. However, this method is approximate and does not provide confidence intervals.

Models have been developed for four different types of income. In some cases slight inconsistencies (when examining point estimates) may occur between the income types for particular MSOAs, e.g. a MSOA may have a larger modelled estimate for net weekly household income (unequivalised) when compared with total weekly household income (unequivalised). Although there may be some such inconsistencies the models selected are the best possible to model the general patterns of income over all MSOAs. This reinforces the need to look at the confidence intervals for the income estimates not just the point estimate since the confidence intervals summarise the variability in the estimates caused by the modelling process.

The model-based method has been developed to ensure that the model-based estimates for MSOAs are constrained to direct survey estimates from the FRS at the region level in England and the country level for Wales. However, the model-based estimates will not be consistent with FRS estimates of average weekly household income for other geographical levels.

These estimates have been produced on MSOA boundaries. Users must be aware of this when using the estimates in any application or drawing conclusions from the data. The estimates are also based on 2013/14 survey data and so are only valid for this period.

Appendix

A. Model Procedures

A.1 Basic Theory

This appendix covers the standard small area estimation theory used by SAEP to produce estimates of income. Some of the basic theory of Chapter 3 is repeated and then extended to cover all the methodology of the modelling procedure. For more information on the general SAEP modelling procedure refer to the SAEP report Heady et al (2003).

We are interested in estimating the mean (also known as the average) of a survey variable (weekly household income) within a set of small areas. Denoting the survey variable as Y , we want to find

$$\widehat{\bar{Y}}_j \quad [5]$$

where the line above indicates a mean, the hat indicates an estimate and the subscript j indicates the area. For example, if Y was weekly household income and we wanted MSOA-level estimates, then $\widehat{\bar{Y}}_j$ would be the estimate of mean weekly household income in MSOA j .

A.2 Basic Theory

A relationship (model) is assumed between the survey variable, Y , and the covariate, X (at present we will only assume one covariate, denoted as X , for simplicity). Due to confidentiality constraints in the UK and the availability of data it is not possible to include individual or household level covariates in the model, i.e. age, sex etc. So this means the covariates are limited to those at an area level. So we write the basic multilevel model as:

$$y_{ij} = \alpha + \beta \bar{X}_j + u_j + e_{ij} \quad [6]$$

where:

y_{ij} is the survey variable of interest for individual/case i within area of interest j

\bar{X}_j is the (known) population j mean for the covariate in area of interest j (strictly speaking this refers to just one covariate);

α and β are the regression parameters for intercept and slope respectively;

u_j is the area level residual, which is included to account for area level means to randomly differ from the fixed part (the overall trend), $\alpha + \beta \bar{X}_j$, of the model, and is assumed to have expectation 0 and variance σ_u^2 ; and

e_{ij} is the individual “within area” residual, with expectation 0 and variance σ_e^2 .

We denote the estimates of the parameter values as $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ (the details of the methods used to obtain these are described later).

The general estimator tool that has been used in the SAEP for UK data is the common SYNTH estimator:

$$\hat{Y}_{j,SYNTH} = \hat{\alpha} + \hat{\beta}\bar{X}_j. \quad [7]$$

The SYNTH uses the (fixed part) area estimate from the model in Equation [6] as the final small area estimate. Other estimates exist such as the Generalised Regression (GREG) and Composite (COMP) estimator. However these are not appropriate for this application since sample data is not available for all small areas.

A.3 General SAEP Theory

The model in Equation [6] needs to be modified due to a further complication concerned with survey design. Samples are drawn using postcode sectors (PCS) as the primary sampling unit (PSU) - forcing random variation to be modelled using PCS as the area level j . However, area estimates are required for MSOAs so covariates are required for these MSOAs and not PCS.

We adapt Equation [6] to allow for these facts to the following:

$$y_{ij} = \alpha + \beta\bar{X}_{k(ij)} + u_j + e_{ij} \quad [8]$$

where $k_{(ij)}$ (generally denoted just 'k' in this report) relates to the MSOA that unit i in PCS j falls within. Hence, the residuals refer to between and within PCS variation, but the resulting estimates will be at the covariate area level.

A.3.1 Calculation of Confidence Intervals

As well as producing estimates of the mean within the small area, it is also important to be able to assess the accuracy of these estimates. Under the assumptions of the model the unknown true value of the area mean (\bar{Y}_k) is itself a random variable – with variance σ_u^2 and expectation $(\alpha + \beta\bar{X}_k)$. If we ignore the possibility that a small part of the sample data may have come from area k , $\hat{Y}_{k,SYNTH}$, the SYNTH estimator can be treated as an independent random variable with the same expectation. In order to set confidence intervals we need to calculate the variance of the difference between these two random variables – i.e. $V(\hat{Y}_{k,SYNTH} - \bar{Y}_k)$. This quantity, which is known as the Mean Square Error (MSE) of $\hat{Y}_{k,SYNTH}$, can be divided into two elements as shown in Equation [5].

$$MSE(\hat{Y}_{k,SYNTH}) = V(\bar{Y}_k) + V(\hat{Y}_{k,SYNTH}) = V(\bar{Y}_k) + E\left[\left((\hat{\alpha} + \hat{\beta}\bar{X}_k) - (\alpha + \beta\bar{X}_k)\right)^2\right] \quad [9]$$

Re-expressing the terms on the right hand side of the equation gives us:

$$MSE(\hat{Y}_{k,SYNTH}) = \sigma_u^2 + V(\hat{\alpha}) + 2Covariance(\hat{\alpha}, \hat{\beta})\bar{X}_k + V(\hat{\beta})\bar{X}_k^2 \quad [10]$$

An estimate of the MSE can be obtained by substituting estimated parameter values (e.g. $\hat{\sigma}_u^2$ for σ_u^2) in the above formula.

However, this formula cannot be applied directly in our case. The formula requires that $\hat{\sigma}_u^2$ is the “between estimation area” variance, but our $\hat{\sigma}_u^2$ relates to PCS, which are not our estimation areas. To enable calculation of confidence intervals we have made the following assumption.

A.3.2 Variance Assumption

We assume that because MSOAs and PCS are of similar size in terms of households the variation for MSOAs will be of a similar size to that for PCS and so the σ_u^2 from Equation [10] is similar for both kinds of area. This assumption allows us to use the value of $\hat{\sigma}_u^2$ derived from the model given in Equation [10] in error calculations relating to MSOAs.

So, given this assumption, a 95% confidence interval for the area model estimate for MSOA k , is:

$$\hat{\alpha} + \hat{\beta}\bar{X}_k \pm 1.96(\hat{\sigma}_u^2 + \underline{X}^T \text{Var}(\hat{\beta})\underline{X})^{1/2} \quad [11]$$

where:

\underline{X} represents the vector, for the MSOA concerned, of the values of the covariates in the model (including in this case the one of value unity representing the intercept)

$\hat{\sigma}_u^2$ is the estimated between-PCS variance

$\text{Var}(\hat{\beta})$ is the estimated variance/covariance matrix of the parameter estimates.

Note that initial investigations have shown that this assumption may result in confidence intervals being conservative.

A.4 Small Area Estimation (SAEP) Income Model

We now apply the general SAEP theory to the particular case of modelling income.

For all four types of income modelled in this project the variable is not normally distributed but positively skewed (the largest values differ from the mean more than the smaller values do). By using the natural logarithm (ln) of the appropriate type of income as the response variable this skewness is reduced and it is assumed for the analysis that the transformed variable follows a normal distribution. Therefore the model used here, known as the lognormal, for modelling income is;

$$\ln(y_{ij}) = \alpha + \beta\bar{X}_{k(ij)} + u_j + e_{ij} \quad [12]$$

y_{ij} is weekly income for household i PCS j ;

$\bar{X}_{k(ij)}$ is the population mean for the covariate in MSOA k that household i in PCS j falls within;

α and β are the regression parameters for intercept and slope respectively;

u_j is the area level residual assumed to have expectation 0 and variance σ_u^2 ; and

e_{ij} is the individual within area residual, with expectation 0 and variance σ_e^2 .

So estimates produced from this model will be on the ln scale. In order to obtain small area estimates on the original scale we use the back transformation of ln, the exponential. So the final estimate of mean household weekly income for MSOA k is:

$$\hat{Y}_{k,SYNTH} = \exp\left(\hat{\alpha} + \hat{\beta}\bar{X}_k + \frac{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}{2}\right). \quad [13]$$

The extra term inside the brackets in Equation [13], $\frac{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}{2}$, is the bias correction factor. This adjusts for the bias due to applying the back transformation.

Since the model involves transformations, the confidence intervals will be specified in terms of the transformed scale. Confidence intervals in the original scale can be obtained by applying back-transformations to the upper and lower confidence limits.

In the calculation of some diagnostic tests (see Chapter 6) the standard error for $\hat{Y}_{k,SYNTH}$ is also required. An approximate method is employed to calculate these standard errors. Let $CI_{\min k}$ and $CI_{\max k}$ denote the upper and lower limits of the 95% confidence interval for $\hat{Y}_{k,SYNTH}$ on the original scale (after the back-transformation). The standard error of $\hat{Y}_{k,SYNTH}$ is defined as follows:

$$s.e.(\hat{Y}_{k,SYNTH}) = \frac{\max(CI_{\max k} - \hat{Y}_{k,SYNTH}, \hat{Y}_{k,SYNTH} - CI_{\min k})}{1.96}. \quad [14]$$

A.5 Adding auxiliary data to the model

In order to select the covariates (auxiliary variables) to be included in each model a stepwise forward selection process has been used. The initial stage of modelling is carried out in SAS. The model developed in SAS does not take into account the multilevel structure of the data, however previous investigations show that those covariates considered to be significant in a single level model are also significant in a multilevel model. During the modelling process the following are taken into account:

- All regional covariates were forced into the model
- Logit transformations of covariates that are proportions were considered for inclusion in the modelling
- Interactions of significant covariates already included in the model were also considered for inclusion in the final model

A selection criterion based on the T statistic has been used. Covariates were considered as being significant if absolute value of T was greater than 2. Note some covariates have been included in the model even though they are not considered to be significant using the T rule since they are included in an interaction term which is significant.

After this modelling process is carried out the final model covariates are entered into SAS to obtain estimates of the parameters when a multilevel model structure is used.

A.6 Benchmarking

After modelling the method of benchmarking is used to adjust the estimates to avoid inconsistencies between the model estimates and direct survey estimates. The FRS survey data are used to calculate direct

estimates of income at the Region and country level for England and Wales respectively (Shale et al (2015)). Model-based MSOA estimates of income are aggregated to the Region level in England and country level in Wales and comparisons made between the two sets of estimates. The ratio of direct survey estimate to aggregated model estimate at the Region/country level is used to scale all model MSOA level estimates and their confidence intervals.

In order to aggregate estimates to a regional level it was necessary to obtain estimates of the number of households per MSOA since:

$$I_R = \frac{\sum_{M \in R} (H_M * I_M)}{H_R}, \quad [15]$$

where,

i_i – estimate of mean income for geographical area i

H_i – number of households for geographical area i

R – region

M – MSOA.

Estimates of the number of households in a MSOA were available from the 2011 Census.

B Data Sources

B.1 Survey Data Income - Definitions

This appendix contains details on the four income types modelled. For more specific information please refer to the survey reports (Shale et al (2015) and DWP (2013)).

B.2 Total household weekly income (unequivalised)

Total household weekly income is the sum of the gross income of every member of the household plus any income from benefits such as Working Families Tax Credit. It is calculated as the sum of income from:

- wages and salaries (gross)
- self-employment
- investments
- tax credits
- state pension and income support/pension credit
- other pensions
- other benefits
- disability benefits
- other sources of income

B.3 Net household weekly income (unequalised)

Net household weekly income (unequalised) is the sum of the net income of every member of the household. It is calculated using the same components as total income but income is net of:

- Income tax payments
- National insurance contributions
- Domestic rates/council tax
- Contributions to occupational pension schemes
- All maintenance and child support payments, which are deducted from the income of the person making the payments
- Parental contribution to students living away from home

B.4 Net household weekly income before housing costs (equalised)

Net household weekly income before housing costs (equalised) is composed of the same elements as net household weekly income but is subject to the OECD's equalisation scale (DWP 2013)). Note that net household weekly income was previously subject to the McClement's equalisation scale.

Applying an equalisation scale adjusts the household income values to take into consideration the number and composition of people in the household; it represents the income level of every individual in the household. Equalisation is needed in order to make sensible income comparisons between households. For example, one household may have 2 adults and 2 children and have a total weekly household income of £300. If this is compared with a household containing just 1 adult who has a total weekly household income of £270, then although the first household has the higher total weekly income it is the second that has the higher standard of living.

Although a number of equalisation scales have been developed, the equalisation scale used for the income estimates is the OECD's scale. An example of the effect of applying the OECD's scale is as follows:

A single person, a couple and a couple with two children aged four and seven, all have unequalised net weekly household incomes of £100 before housing costs. After equalisation, these become £164 (single person); £100 (couple); £72 (couple with children).

B.5 Net household weekly income after housing costs (equalised)

Net household weekly income after housing costs (equalised) is composed of the same elements of net household weekly income but is subject to the following deductions prior to the OECD's equalisation scale being applied:

- Rent (gross of housing benefit)
- Water rates, community water charges and council water charges
- Mortgage interest payments (net of any tax relief)
- Structural insurance premiums (for owner occupiers)
- Ground rent and service charges

B.6 FRS and Households Below Average Income (HBAI) Data

All of the survey data used in the modelling process are obtained from the FRS. However two of the income types above are defined by a different study that is based on FRS data. Net weekly household income (equalised) both before and after housing costs is defined and calculated in the HBAI report (DWP 2013)). Although all four types of income for a particular household will be calculated using the same FRS data the

HBAI methodology makes some changes to the original data set. The HBAI data set is a cut down version of the FRS data since the HBAI excludes households containing a married adult whose spouse is temporarily absent. An adjustment is also made to sample cases at the top of the income distribution to correct for volatility in the highest income captured in the survey. For more detail on these adjustments and the reasons for them see the HBAI documentation (DWP (2013)). Note that due to the differences in the HBAI and FRS methodology the two sets of data have different grossing factors.

B.7 Auxiliary Data Sources and Covariates

This appendix contains specific details on each of the data sources including the population estimates used to produce the models for England & Wales. More information on the specific variables obtained from the data sources are given with any appropriate technical detail. All variables were obtained or derived to a MSA-level. The auxiliary data sets considered for inclusion in modelling income are listed below.

- Census, 2011
- Department for Work and Pensions benefit claimant counts, August 2013
- Valuation Office Agency Council Tax Bandings, March 2013
- Her Majesty's Revenue and Customs, Child Tax Credit and Working Tax Credit, Aug 2013
- Office for National Statistics, House Price Statistics for Small Areas, year ending March 2014
- Department of Energy & Climate Change, Energy Consumption data, 2013
- Regional/country identification variable

The DWP data were provided as counts. However it was more appropriate to include proportions or prevalence rates in the modelling process. MSA population data were used as denominators to derive these proportions.

Covariates were centred by subtracting the corresponding means for England and Wales. Centring the covariates enables easier interpretation of the model parameters, e.g. the intercept now represents the weighted mean over all areas of the response variable (after the log transformation). Covariates were considered for inclusion in the model on the original as well as the transformed logit scale.

The model selection process for the 2013/14 small area income estimates used variables that were relevant to the time period, so some of the DWP and HMRC variables in Tables 13 and 15 are calculated from the benefits data that were available in 2013/14. The following benefits from these tables have since been replaced with other benefits:

- Incapacity Benefit has been replaced by Employment and Support Allowance
- Disability Living Allowance has been replaced by Personal Independence Payment and Attendance Allowance
- Income Support, Income-related Employment and Support Allowance, Income-based Jobseekers Allowance, Child Tax Credit and Working Tax Credit are being replaced by Universal Credit

B.7.1 Census Data 2011

The following Census variables were considered for inclusion in modelling income.

Table 12: Variables considered for inclusion in modelling income, Census 2011

Variable name	Label
phouse	Proportion of household spaces that are detached, semi detached or terraced
pflat	Percentage of household spaces that are a flat, maisonette or commercial building
pchbath	Proportion of households with sole use of a bath/shower and toilet and central heating
p12rooms	Proportion of households with one or two rooms
avhhpeop	Average number of people per household
avhhroom	Average number of rooms per household
pgroupab	Proportion of people aged 16 to 74 whose approximated social grade is AB
pgroupc1	Proportion of people aged 16 to 74 whose approximated social grade is C1
pgroupc2	Proportion of people aged 16 to 74 whose approximated social grade is C2
pgroupd	Proportion of people aged 16 to 74 whose approximated social grade is D
pgroupe	Proportion of people aged 16 to 74 whose approximated social grade is E
pnocar	Proportion of households that do not have a car or van
ponecar	Proportion of households that have one car or van
pcare	Proportion of people providing unpaid care
pcommun	Proportion of people living in communal establishments
pbornuk	Proportion of people born in the UK
pborneur	Proportion of people born in Europe
phhdepch	Proportion of households with dependent child(ren)
pecactiv	Proportion of people aged 16 to 74 who are economically active
phrpecac	Proportion of household reference persons aged 16 to 74 who are economically active
punemp	Proportion of people aged 16 to 74 who are unemployed
pftstud	Proportion of people aged 16 to 74 who are full-time students
pltunemp	Proportion of people aged 16 to 74 who are long-term unemployed
pemployd	Proportion of people aged 16 to 74 who are employed or self-employed
pretired	Proportion of people aged 16 to 74 who are retired
pnonwbri	Proportion of people who are 'Not White British'
phealth	Proportion of people in households reporting good or fairly good health
phhtype1	Proportion of households that contain one person only
phhtype2	Proportion of households that are lone parent households
phhtype3	Proportion of households that are lone parent with dependent child(ren)
phhtype4	Proportion of households that are lone parent with all child(ren) non - dependent
phhtype5	Proportion of households that are a couple with no children
phhtype6	Proportion of households that are a couple with dependent child(ren)
phhtype7	Proportion of households that are a couple with all child(ren) non -dependent
phhdepr	Proportion of households classed as deprived
pcouple	Proportion of people in households that are living in a couple

phhfloor	Proportion of households whose lowest floor level is the basement or the ground floor
pltli	Proportion of people in households with a long-term limiting illness
pswd	Proportion of people aged over 16 who are single, separated, widowed or divorced
pmanprof	Proportion of people aged 16 to 74 whose NS-SEC is 'managerial and professional'
pintocc	Proportion of people aged 16 to 74 whose NS-SEC is 'intermediate'
proutman	Proportion of people age 16 to 74 whose NS-SEC is 'routine and manual'
phrpman	Proportion of HRPs aged 16 to 74 whose NS-SEC is 'managerial and professional'
phrpint	Proportion of HRPs aged 16 to 74 whose NS-SEC is 'intermediate'
phrprou	Proportion of HRPs aged 16 to 74 whose NS-SEC is 'routine and manual'
povercrw	Proportion of households that are overcrowded
pqual34	Proportion of people aged 16 to 74 whose highest qualification is level 3 and level 4
prelig	Proportion of people who have a religion
phrpreli	Proportion of household reference persons who have a religion
phrpmale	Proportion of household reference persons who are male
phhshare	Proportion of household residents living in a shared dwelling
phhstud	Proportion of households with at least one full-time student or schoolchild living away during term-time
pownocc	Proportion of households that are owner occupied
phhrent	Proportion of households that are rented

B.7.2 DWP Benefit Data 2013

The DWP benefit data obtained were in the format of counts for each benefit type by MSOA. These counts were transformed into proportions using MSOA population estimates, mid-2013.

Table 13 lists the different DWP variables considered for inclusion in the models as well as the population estimate used as a denominator.

Table 13: Variables considered for inclusion in modelling income, DWP benefit claimant counts 2013

Variable name	Label
DLATOTAL	Proportion of people claiming Disability Living Allowance
DCTOTALM	Proportion of males claiming Disability Living Allowance
DCTOTALF	Proportion of females claiming Disability Living Allowance
DCCDLESS12M	Proportion of people claiming Disability Living Allowance with a claim duration of less than 12 months
DCCD1_2YEARS	Proportion of people claiming Disability Living Allowance with a claim duration of 1-2 years
DCCD2_5YEARS	Proportion of people claiming Disability Living Allowance with a claim duration of 2-5 years
DCCD5YEARSOVER	Proportion of people claiming Disability Living Allowance with a claim duration of 5 years and over
DLAMAL	Proportion of people claiming Disability Living Allowance: Mobility Award Lower
DLAMAH	Proportion of people claiming Disability Living Allowance: Mobility Award Higher
DLAMAN	Proportion of people claiming Disability Living Allowance: Mobility Award Nil
DLACAL	Proportion of people claiming Disability Living Allowance: Care Award Lower
DLACAM	Proportion of people claiming Disability Living Allowance: Care Award Middle
DLACAH	Proportion of people claiming Disability Living Allowance: Care Award Higher
DLACAN	Proportion of people claiming Disability Living Allowance: Care Award Nil
PCPTOTAL	Proportion of people aged 60 and over claiming Pension Credit
PCPM	Proportion of males aged 60 and over claiming Pension Credit
PCPF	Proportion of females aged 60 and over claiming Pension Credit
PCPLESS12M	Proportion of people aged 60 and over claiming Pension Credit with a claim duration of less than 12 months
PCP12YEARS	Proportion of people aged 60 and over claiming Pension Credit with a claim duration of 1-2 Years
PCP25YEARS	Proportion of people aged 60 and over claiming Pension Credit with a claim duration of 2-5 years
PCP5YEARPLUS	Proportion of people aged 60 and over claiming Pension Credit with a claim duration of 5 years and over
PCPS	Proportion of single people aged 60 and over claiming Pension Credit
PCPC	Proportion of couples aged 60 and over claiming Pension Credit
PCGEO	Proportion of people aged 60 and over claiming Pension Credit: Guarantee Element Only
PCSEO	Proportion of people aged 65 and over claiming Pension Credit: Saving Element Only

PCGESE	Proportion of people aged 65 and over claiming Pension Credit: Guarantee and Saving Element
IBSDPTOTAL	Proportion of people aged 16 and over claiming Incapacity Benefit/Severe Disablement Allowance
IBSDPMALE	Incapacity Benefit/Severe Disablement Allowance Claimants; Male
IBSDPFEMALE	Incapacity Benefit/Severe Disablement Allowance Claimants; Female
ISPTOTAL	Proportion of people aged 16 and over claiming Income Support
ISPMALE	Proportion of males aged 16 and over claiming Income Support
ISPFEMALE	Proportion of females aged 16 and over claiming Income Support
JSAPTOTAL	Proportion of people aged 16 and over claiming Job Seekers Allowance
JSAMALE	Proportion of males aged 16 and over claiming Job Seekers Allowance
JSAFEMALE	Proportion of females aged 16 and over claiming Job Seekers Allowance

B.7.3 Regional and Country identification variable

England is split into nine regions. Binary variables were created for each region and Wales, taking the value 1 if the MSOA belonged to that region/country and 0 otherwise. These region/country variables are listed below in Table 14. Note that London was selected as the base case and therefore not specified separately in the modelling procedure.

Table 14: Regional variables included in modelling income

Variable name	Country/REGION
northeast	North East
northwst	North West
york	Yorkshire and The Humber
eastmid	East Midlands
westmid	West Midlands
east	East of England
southeast	South East
southwst	South West
wales	Wales

B.7.4 HMRC Child Tax Credit and Working Tax Credit Data 2013

The data were in the form of counts of families or persons receiving a particular type of Tax Credit by MSOA. Counts were centred (but not transformed to the logit scale) and these were tested for inclusion in the models.

Table 15 lists the HMRC variables considered for inclusion in the models.

Table 15: Variables considered for inclusion in modelling income, HMRC Child Tax Credit and Working Tax Credit Data 2013

Variable name	Label
famwktc	Families in Work Receiving; Tax Credit
lpwktc	Lone-Parent Families in Work Receiving; Tax Credit
famwkctwt	Families in Work Receiving; Child Tax Credit and Working Tax Credit
famwkafe	Families in Work Receiving; Child Tax Credit
famwkwt	Families in Work Receiving; Working Tax Credit Only
famwkbfe	Families in Work Receiving; from the Childcare Element
famwkcewtc	Lone-Parent Families in Work Receiving; Childcare Element of Working Tax Credit
famoutct	Families Out of Work Receiving; Child Tax Credit
lpoutct	Lone-Parent Families Out of Work Receiving; Child Tax Credit
cpoutct	Couple Families Out of Work Receiving; Child Tax Credit

B.7.5 Valuation Office Agency council tax band data 2013

Each residential property in England is assigned to one of eight Council Tax bands, depending on its value at 1 April 1991. In Wales, each property is assigned to one of nine Council Tax bands depending on its value at 1 April, 2003. The Council Tax data used here were provided as counts for each band for each MSOA. These counts were transformed into proportions.

The Council Tax bands for England and Wales are not consistent, therefore separate covariates are defined for England and Wales. In Wales, some MSOAs have very high concentrations at one end of the range of tax bands, causing model instability. The final covariates considered for inclusion in the model are as follows:

Table 16: Variables considered for inclusion in modelling income, VOA Council Tax Bands, 2013

Variable name	Label
eabc	Proportion/count of dwellings in English Council Tax bands A, B and C
edef	Proportion/count of dwellings in English Council Tax bands D, E and F
eghi	Proportion/count of dwellings in English Council Tax bands G, H
wabc	Proportion/count of dwellings in Welsh Council Tax bands A, B and C
wdef	Proportion/count of dwellings in Welsh Council Tax bands D, E and F
wghi	Proportion/count of dwellings in Welsh Council Tax bands G, H and I

B.7.6 Department for Energy & Climate Change Energy Consumption data, 2013

Table 17 lists the DECC variables considered for inclusion in the models.

Table 17: Variables considered for inclusion in modelling income, DECC Energy Consumption data 2013

Variable name	Label
ordelecp	Consumption of Ordinary Domestic Electricity as a proportion of total domestic energy consumption
e7elecp	Consumption of Economy 7 Domestic Electricity as a proportion of total domestic energy consumption
gasp	Consumption of Domestic Gas as a proportion of total domestic energy consumption
aordelecc	Average Consumption of Ordinary Domestic Electricity kWh C
ae7elecc	Average Consumption of Economy 7 Domestic Electricity kWh C
agasc	Average Consumption of Domestic Gas kWh C

B.7.7 ONS House Price Statistics for Small Areas, Q1 2014

In addition to counts of the number of dwelling sales, the data contain measures of house prices (e.g. median price) for sales that took place. The data were centred and divided by the standard deviation before being considered for inclusion in the model.

Table 18 lists the HPSSA variables considered for inclusion in the models.

Table 18: Variables considered for inclusion in modelling income, HPSSA 2014

Variable name	Labels
TRNS	Transactions by Dwelling Type; Total Sales
PLQ	Price Indicators for All Dwellings; Lower Quartile
PMED	Price Indicators for All Dwellings; Median
PMEAN	Price Indicators for All Dwellings; Mean

C Data Preparation

Before any modelling could proceed, significant effort had to be channelled into gathering the necessary source data, principally survey response data and covariate data. The survey data set comprises the survey response variables of interest, weekly household income, matched to postcodes, and MSOA codes, for the estimation area. The covariate data set comprises MSOA covariates along with the corresponding MSOA identifiers. These two datasets are matched by reference to the MSOA codes. The resulting matched data set, containing the survey variable along with associated covariates and MSOA and PCS identifiers, becomes the analysis data set. The analysis data set is required for the modelling and the full covariate data set is required to produce the final estimates once the modelling has been performed.

D Results of Modelling for Income

This appendix contains the results of the modelled income estimates for net weekly income (unequalised) and net weekly income (equalised) before and after housing costs. These results are in addition to the

results of the modelling of total weekly income outlined in section 5.1.

Net weekly household income

The estimates of net weekly household income (unequivalised) for MSOAs in England and Wales were produced using the right hand side of Equation [2] in Section 4.3.2 (excluding the u_j and e_{ij} terms).

Figure 14: Model MSOA estimates and 95% Confidence Intervals for net income (unequivalised)

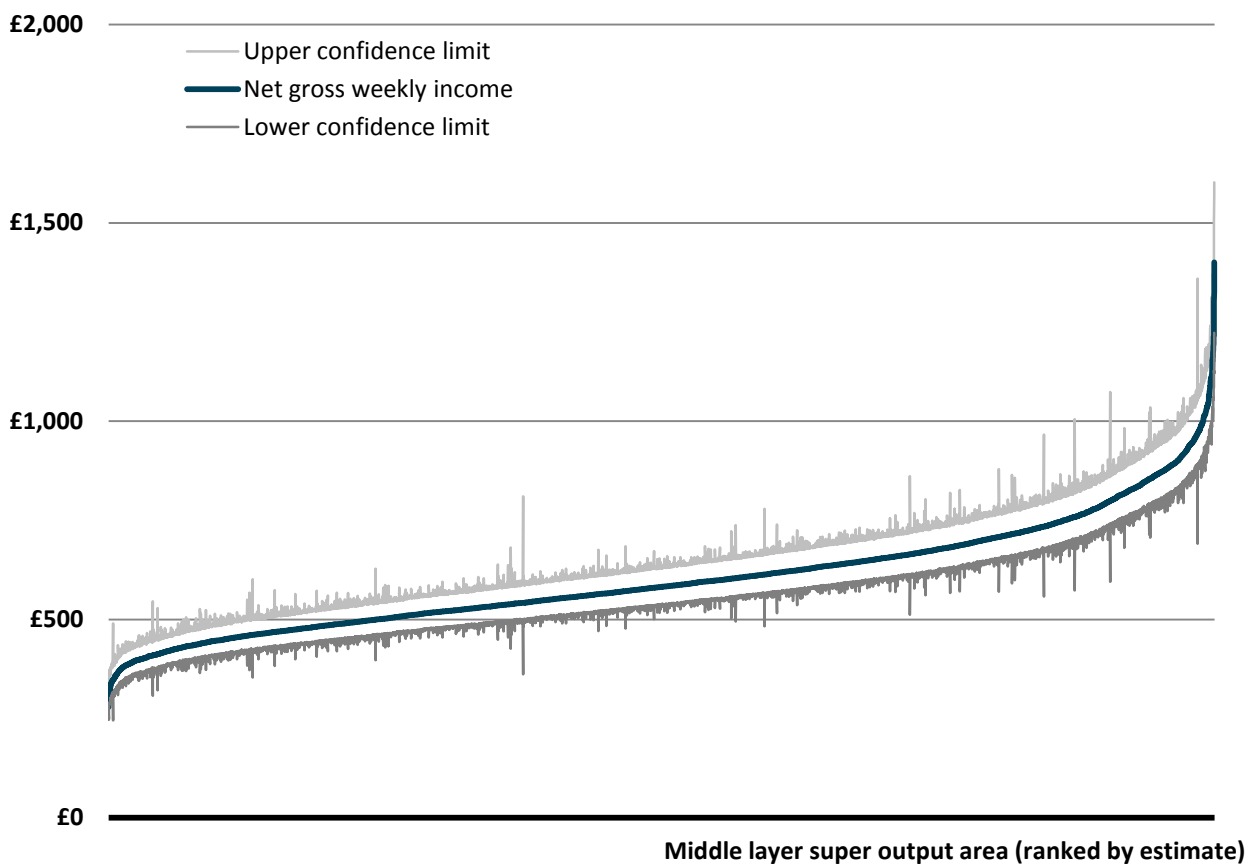
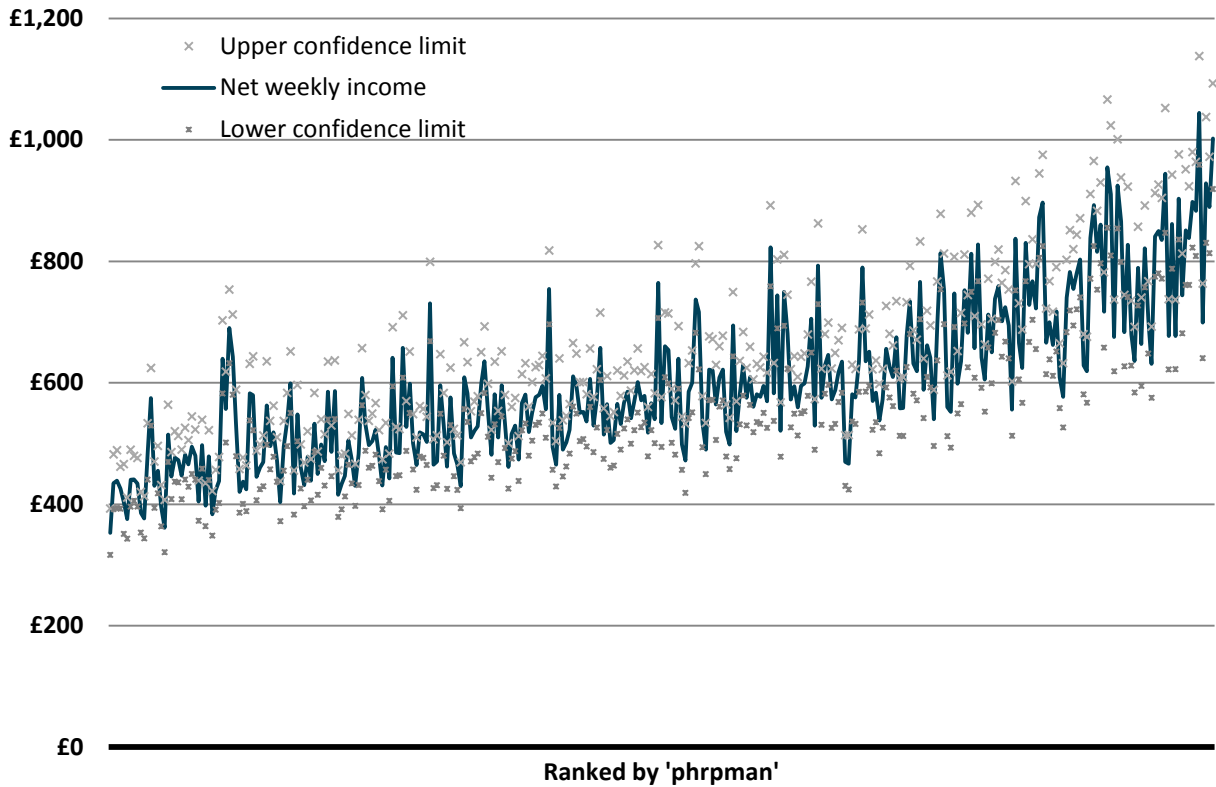


Figure 14 provides a visualisation of the model-based estimates and their 95% confidence intervals. Figure 15 displays a random sample (about 5%) of the MSOA estimates and confidence intervals ranked by the Census covariate 'Phrman' proportion of people aged 16-74 whose NS-SEC is managerial and professional.

Figure 15: Model-based MSOA estimates and 95% Confidence Intervals for net income (unequivalised)



Net Weekly Household Income – Equivalised, Before Housing Costs

The estimates of net weekly household income (equivalised) before housing costs for MSOAs in England and Wales were produced using the right hand side of Equation [3] in Section 4.3.3 (excluding the u_j and e_{ij} terms).

Figure 16 and Figure 17 provide visualisations of the model-based estimates and their 95% confidence intervals

Figure 16: Model-based MSOA estimates and 95% Confidence Intervals for net income, equivalised, before housing costs

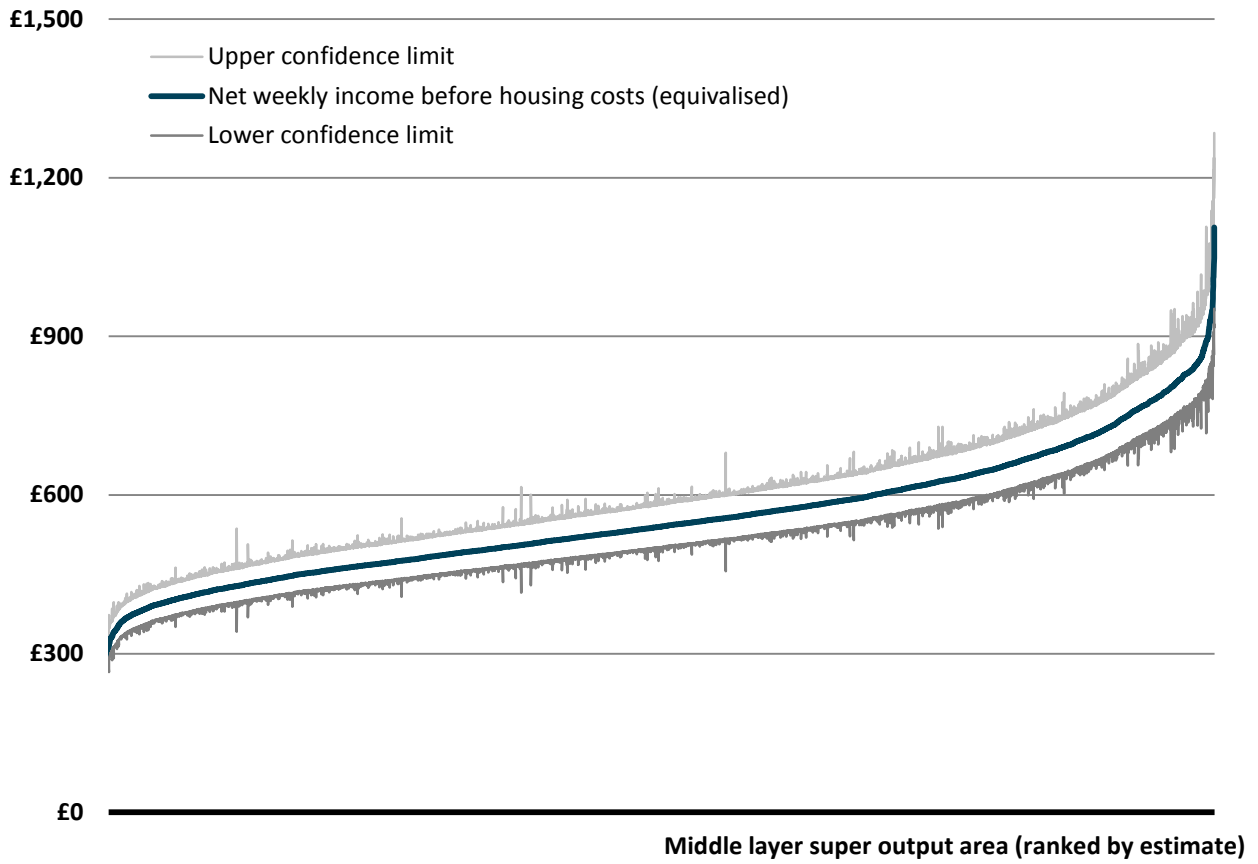
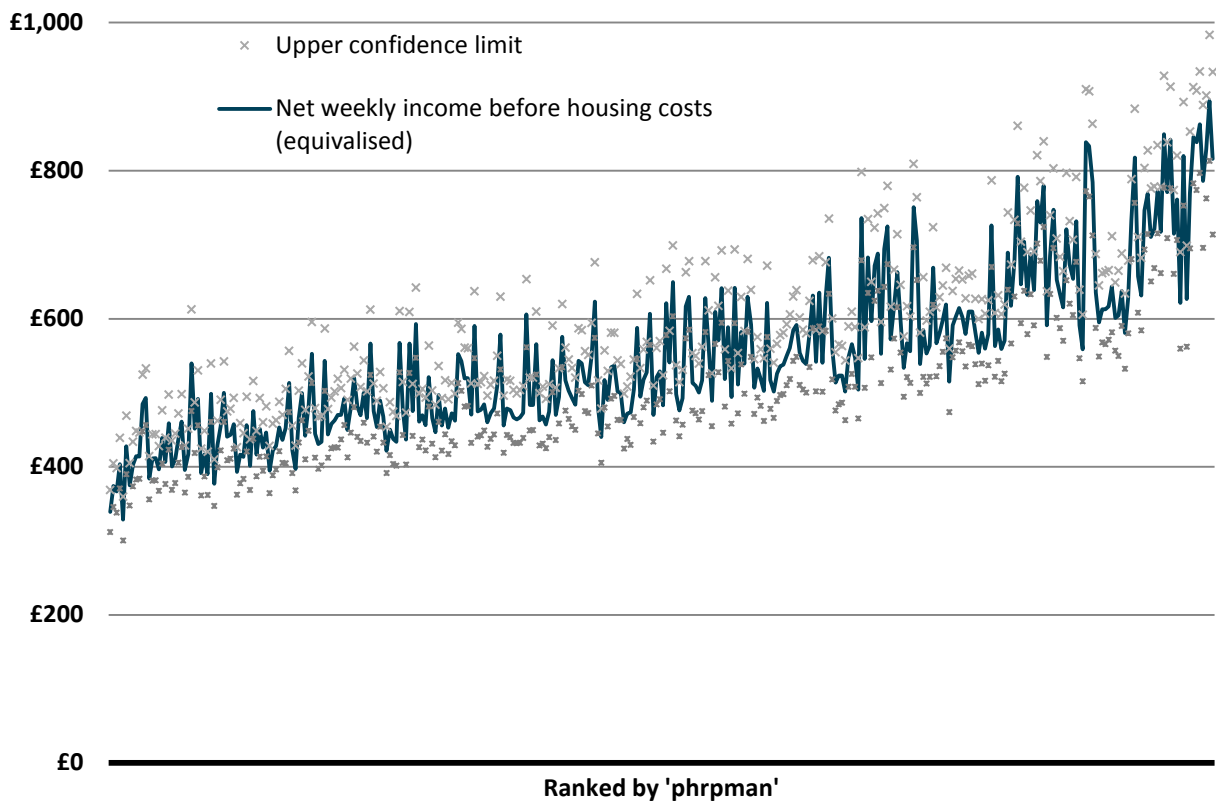


Figure 17: Sample of model-based MSOA estimates and 95% Confidence Intervals for net income, equivalised, before housing costs



Net Weekly Household Income – Equivalised, After Housing Costs

Figure 18 and Figure 19 provide visualisations of the model-based estimates of net income, equivalised after housing costs and their 95% confidence intervals.

Figure 18: Model-based MSOA estimates and 95% Confidence Intervals for net income, equivalised, after housing costs

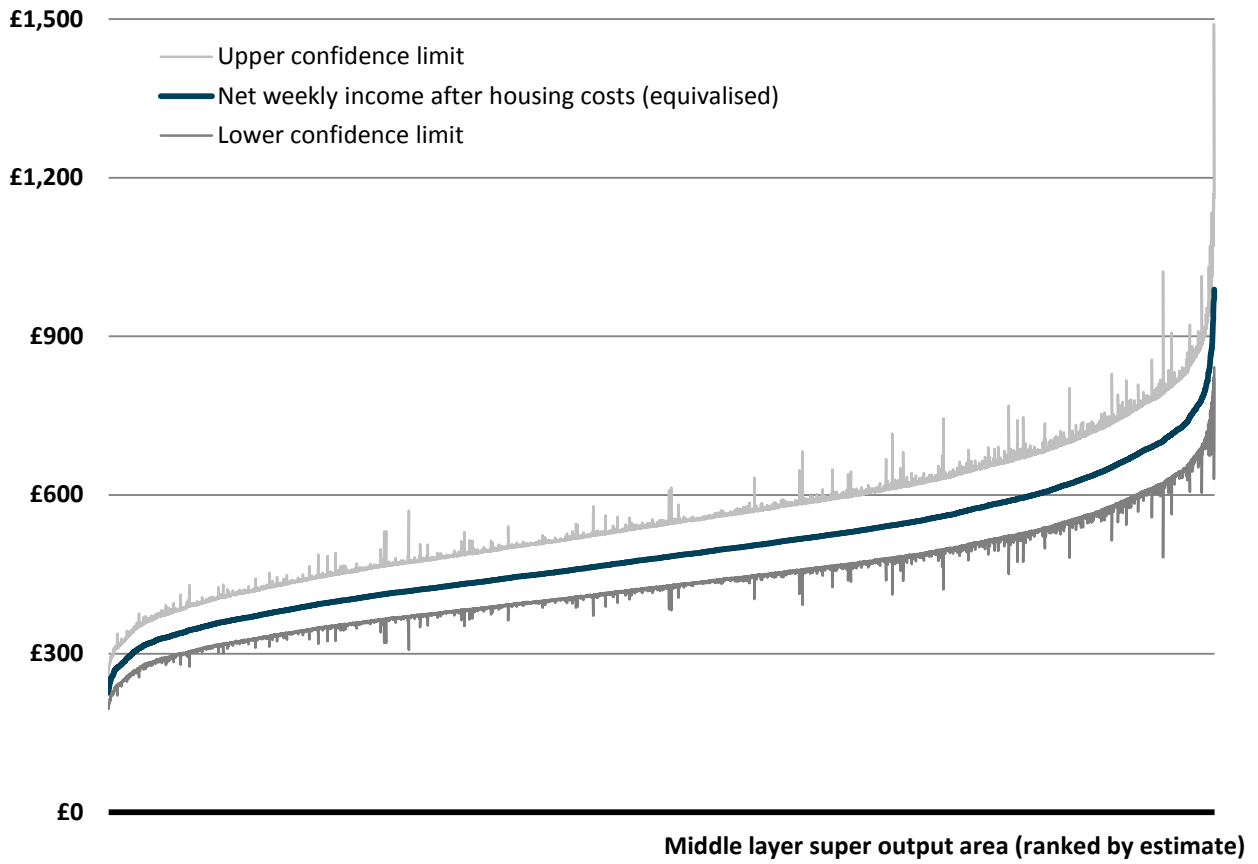
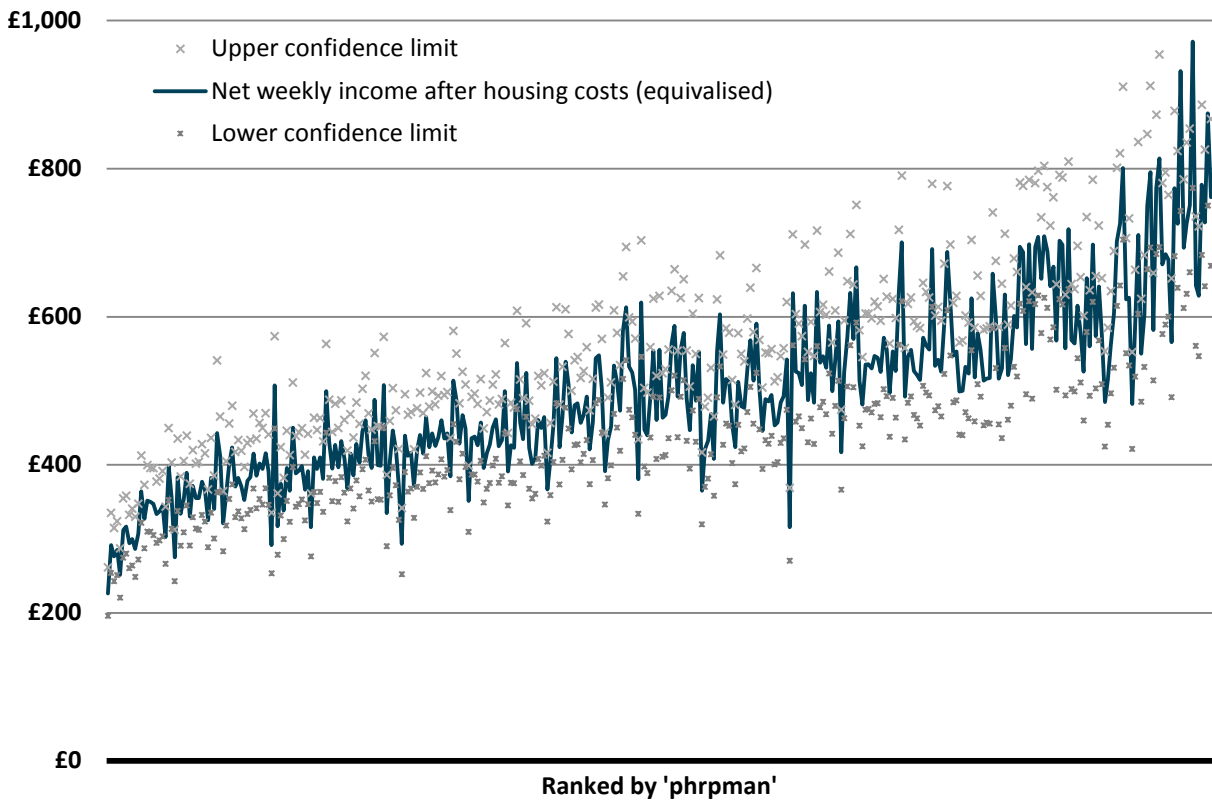


Figure 19: Sample of model-based MSOA estimates and 95% Confidence Intervals for net income, equivalised, after housing costs



E Diagnostic Results

This appendix contains a full set of diagnostic plots for each of the income types modelled to support the results documented in Chapter 6.

Total weekly household income (unequalised)

Figure 20: Household level residuals against model-based estimates, total income (unequalised)

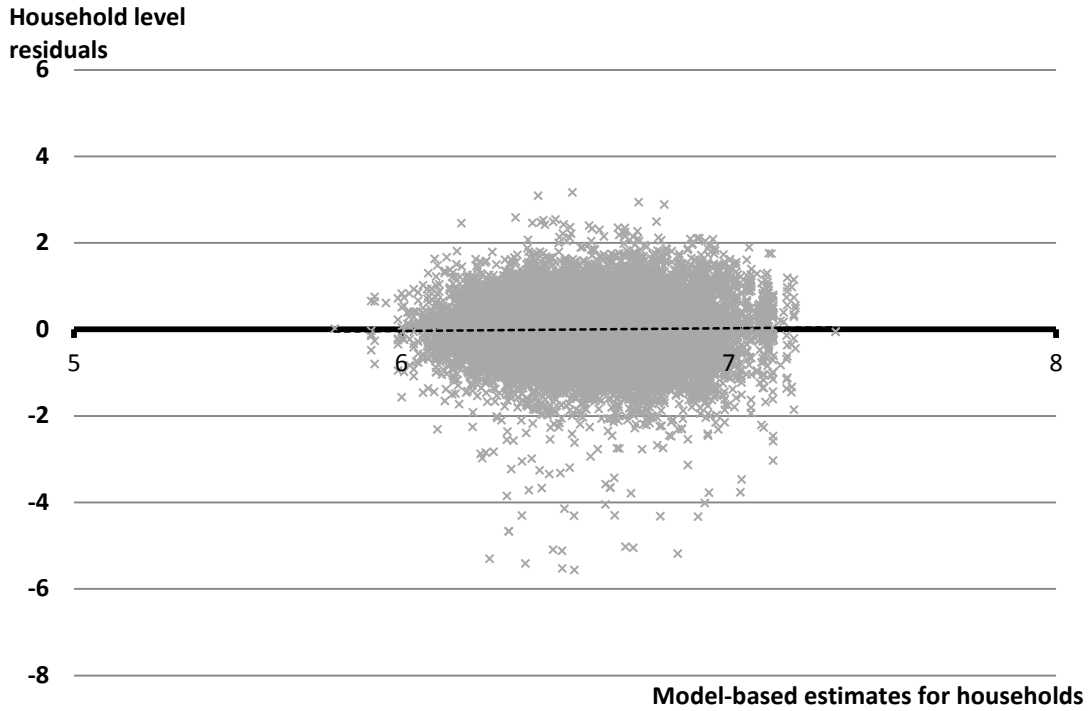


Figure 21: Area level residuals against model estimates, total income (unequalised)

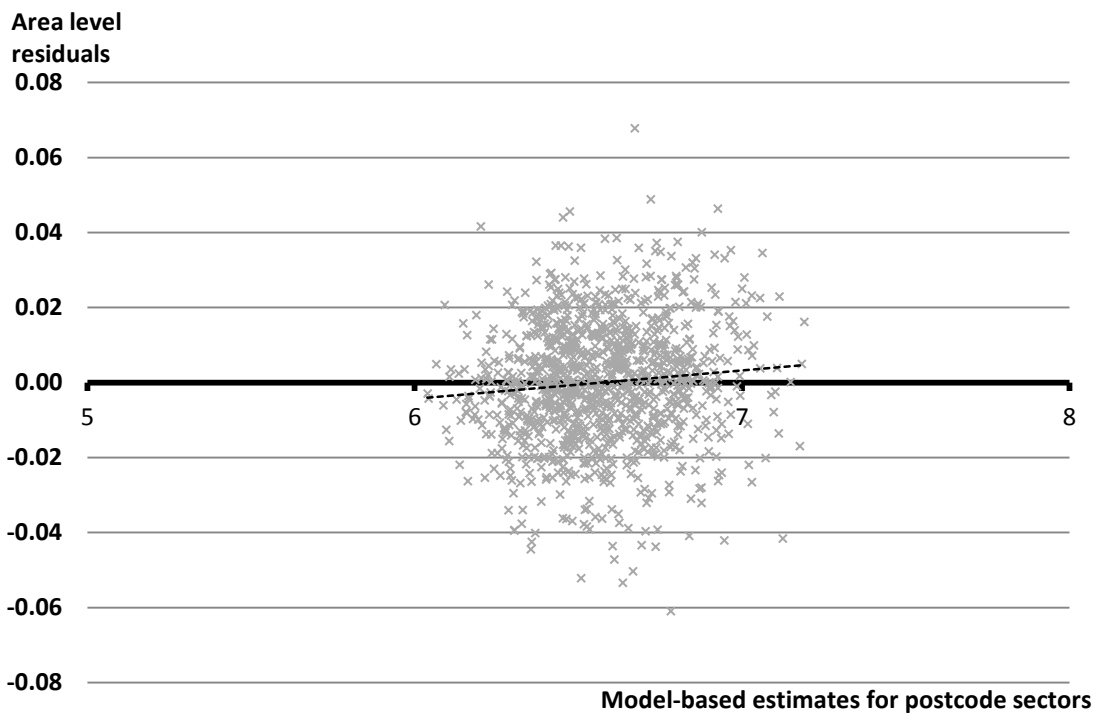


Figure 22: Model-based estimates vs. sample estimates, total income (unequalised)

Direct survey estimates

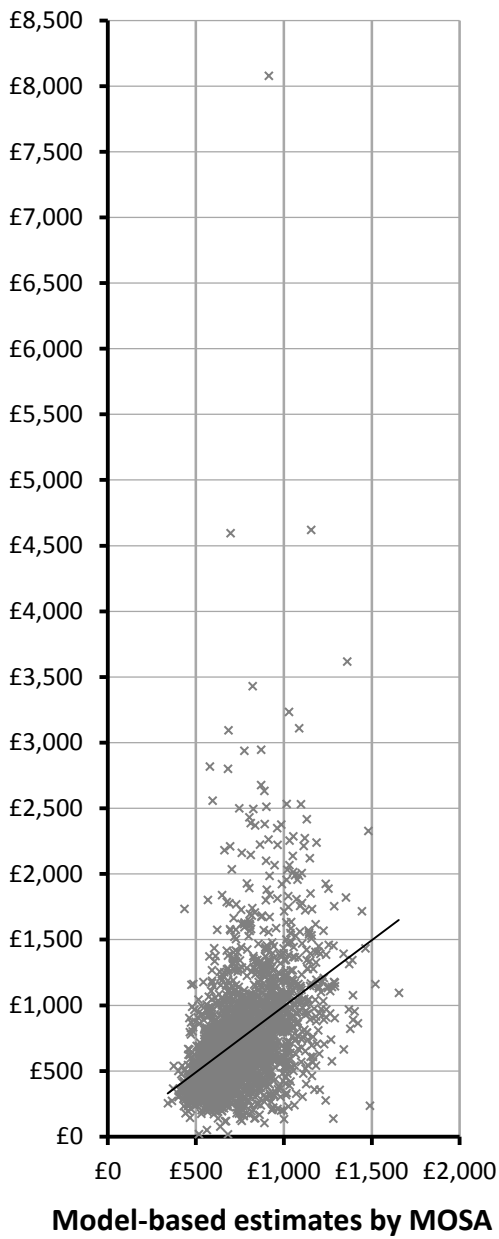
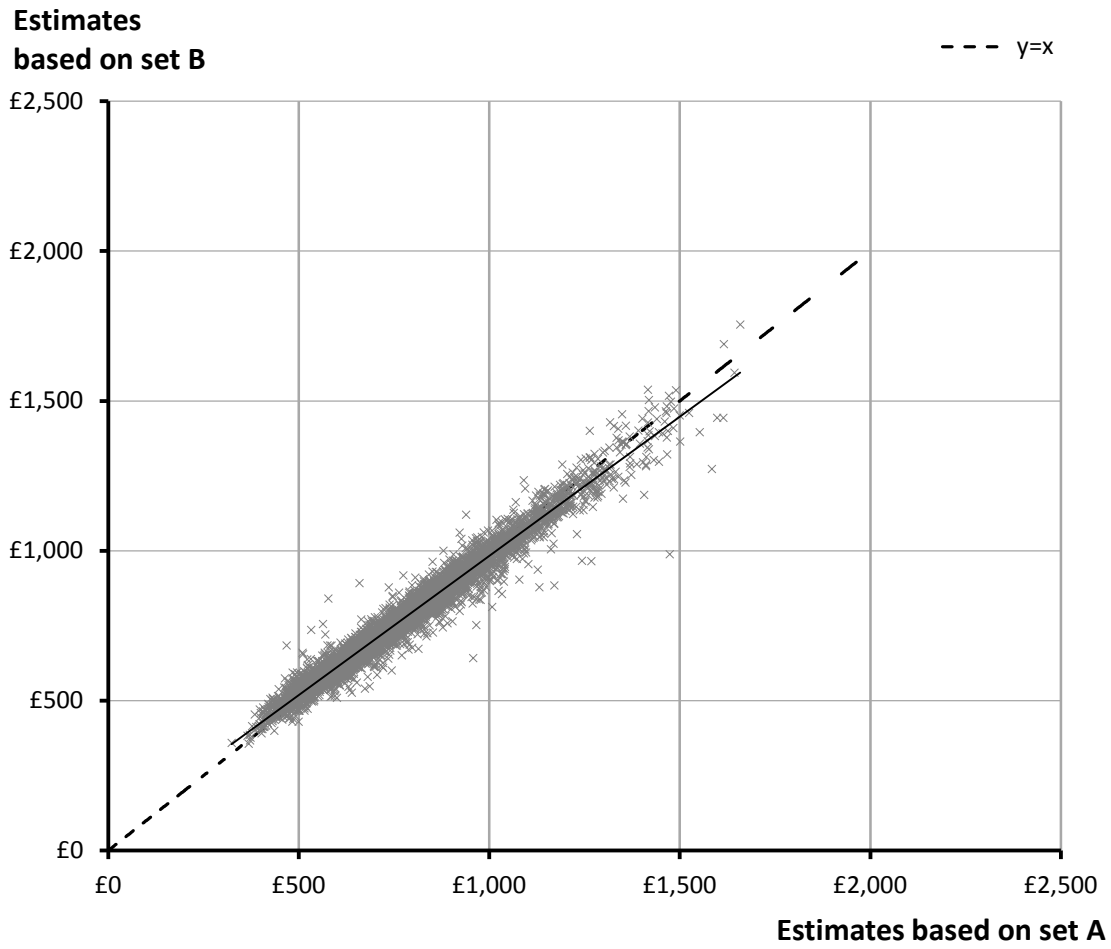


Figure 23: Model estimates from stability analysis, total income (unequivalised)



Map 3 (left): Coefficient of variation - Net weekly income by MSOA, England and Wales, 2013/14;
Map 4 (right): Distance further from estimate of net weekly income of the upper confidence limit than the lower confidence limit (expressed as a percentage of the estimate) by MSOA, England and Wales, 2013/14



Source: Office for National Statistics and Ordnance Survey under the Open Government Licence v3.0.
 Contains OS data © Crown copyright 2016

Figure 24: Household level residuals against model-based estimates, net income (unequalised)

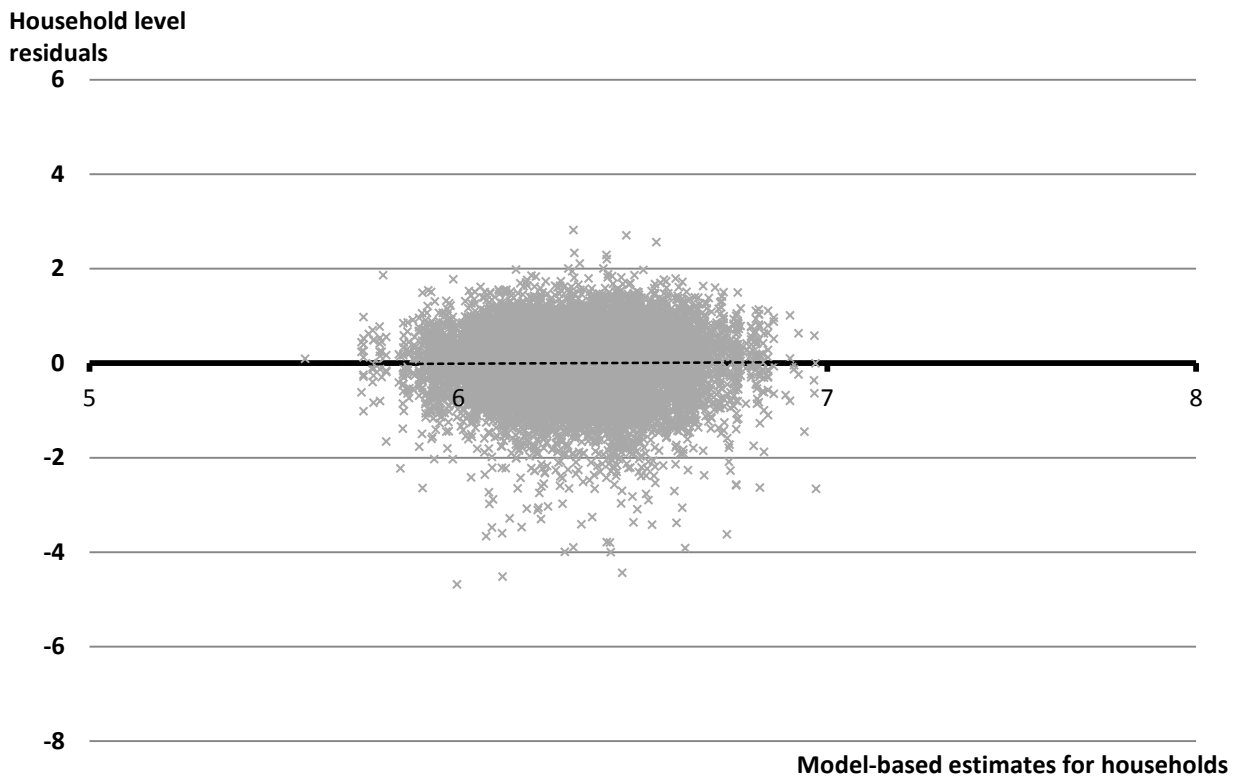


Figure 25: Area level residuals against model-based estimates, net income (unequivalised)

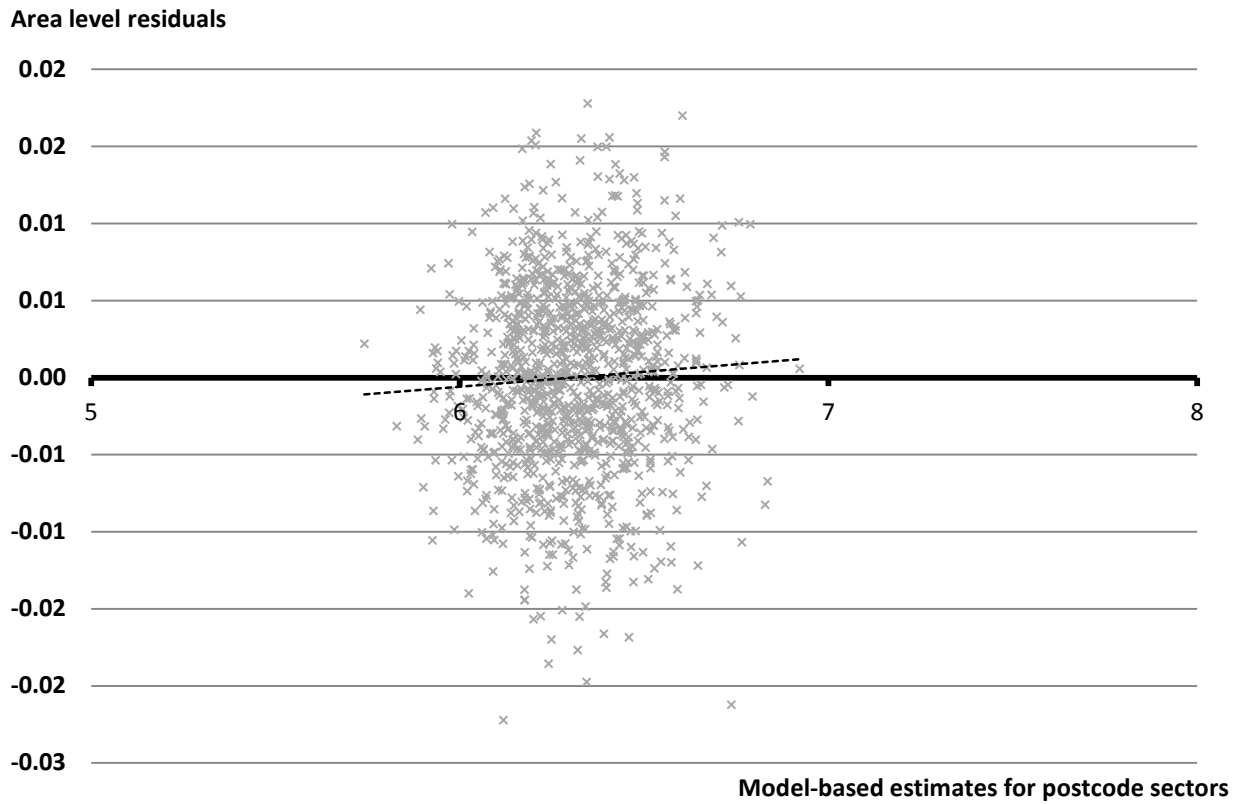


Figure 26: Model-based estimates vs. sample estimates, net income (unequalised)

Direct survey estimates

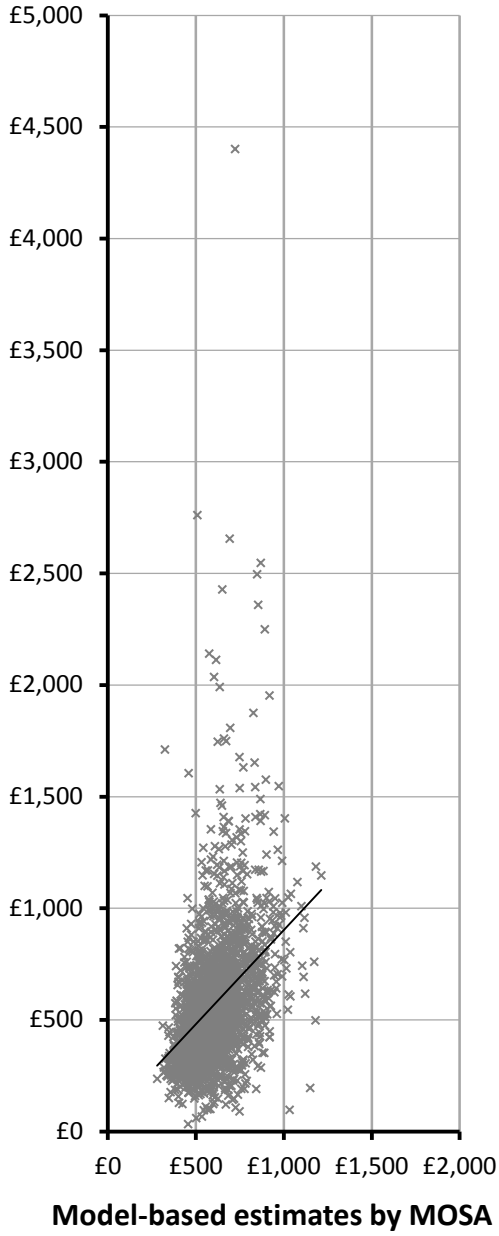
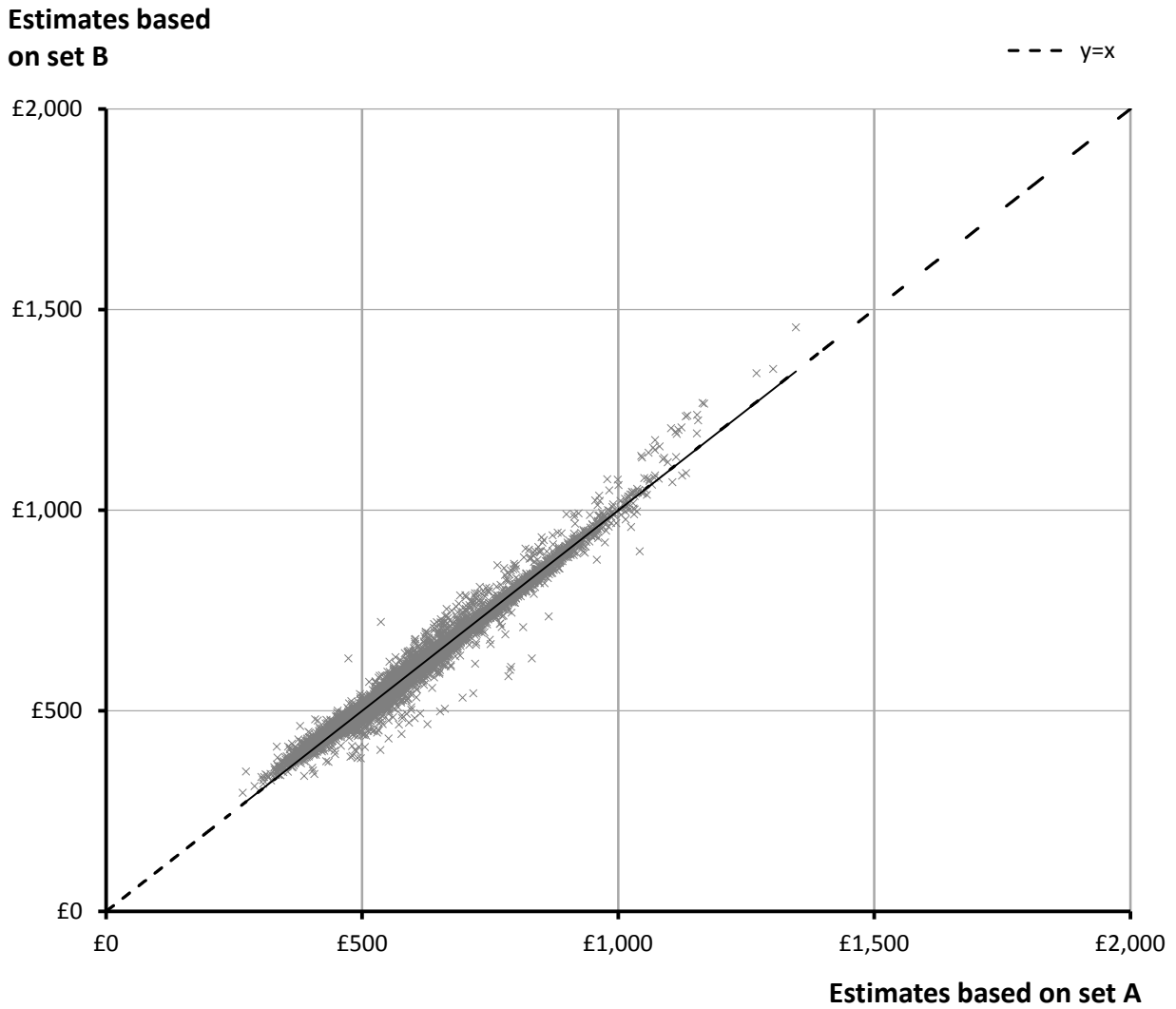
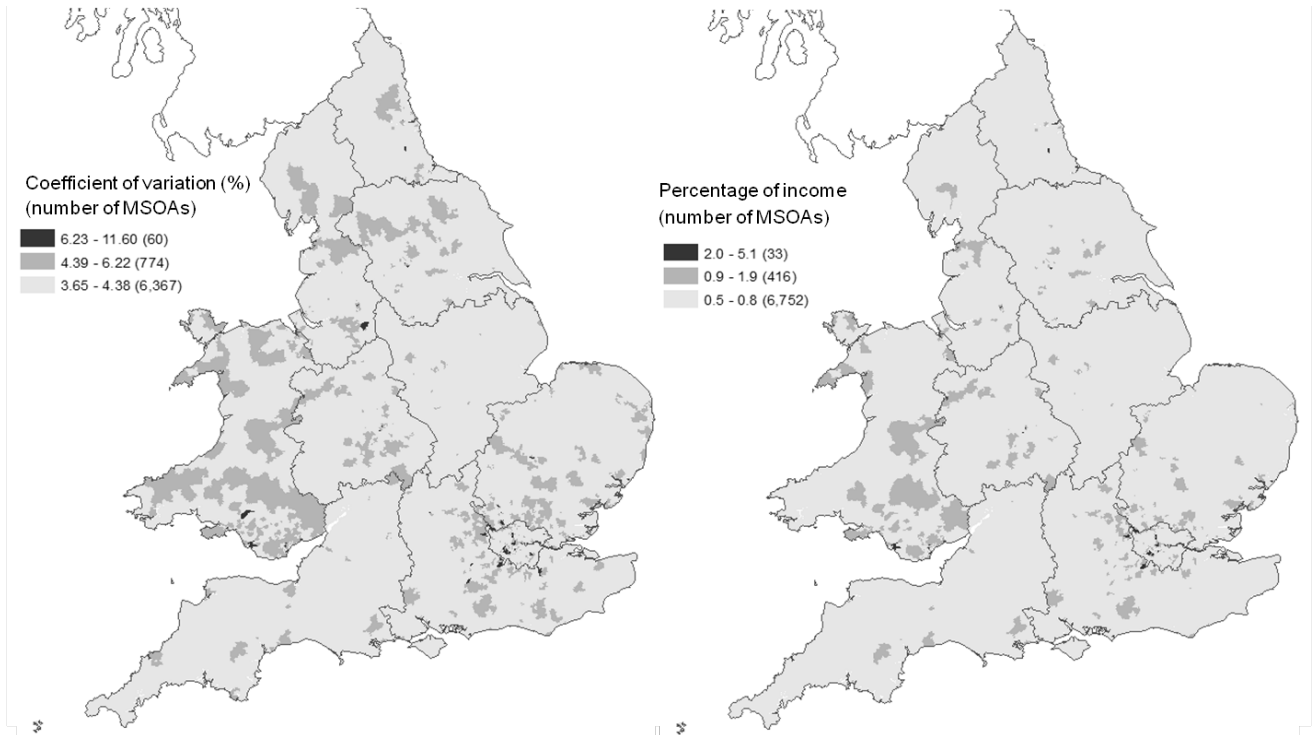


Figure 27: Model-based estimates from stability analysis, net income (unequivalised)



Map 5 (left): Coefficient of variation - Net weekly income before housing costs (equivalised) by MSOA, England and Wales, 2013/14;

Map 6 (right): Distance further from estimate of net weekly income before housing costs (equivalised) of the upper confidence limit than the lower confidence limit (expressed as a percentage of the estimate) by MSOA, England and Wales, 2013/14



Source: Office for National Statistics and Ordnance Survey under the Open Government Licence v3.0. Contains OS data © Crown copyright 2016

Net Weekly Household Income – Equivalised, Before Housing Costs

Figure 28: Household level residuals against model-based estimates, net income, equivalised before housing costs

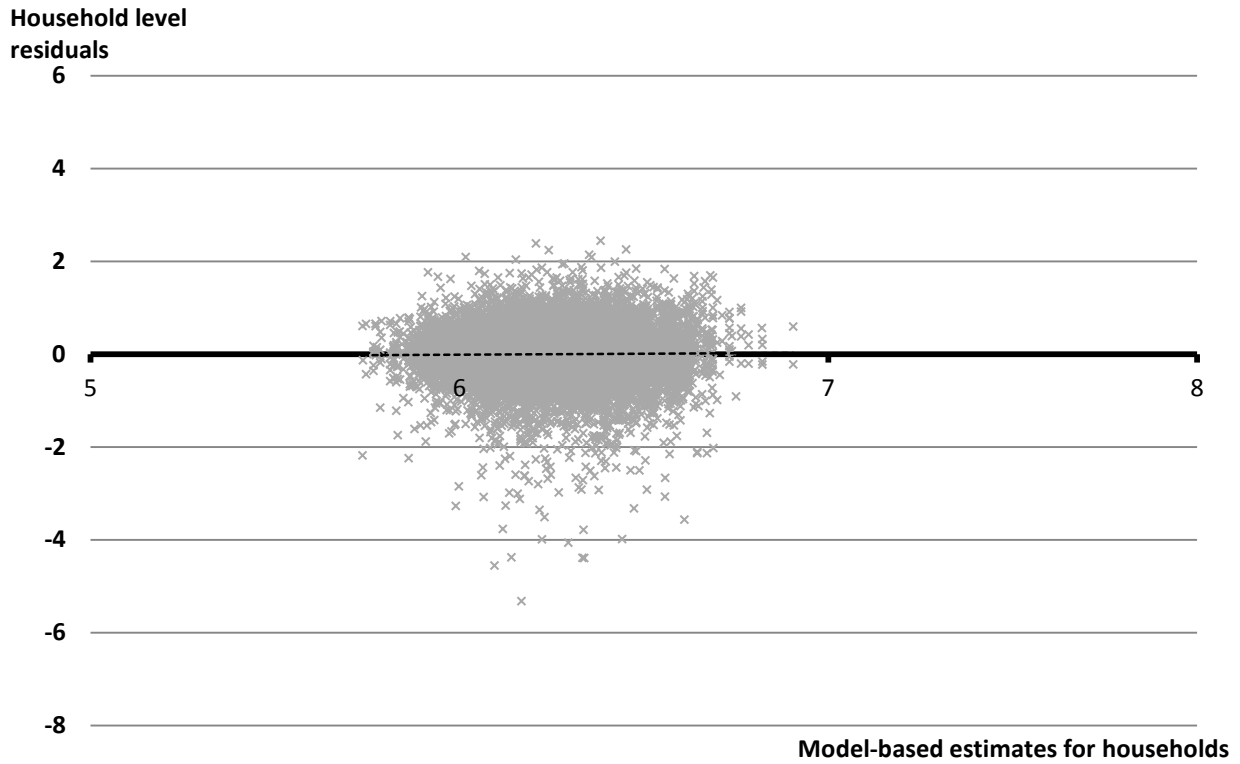


Figure 29: Area level residuals against model-based estimates, net income, equivalised before housing costs

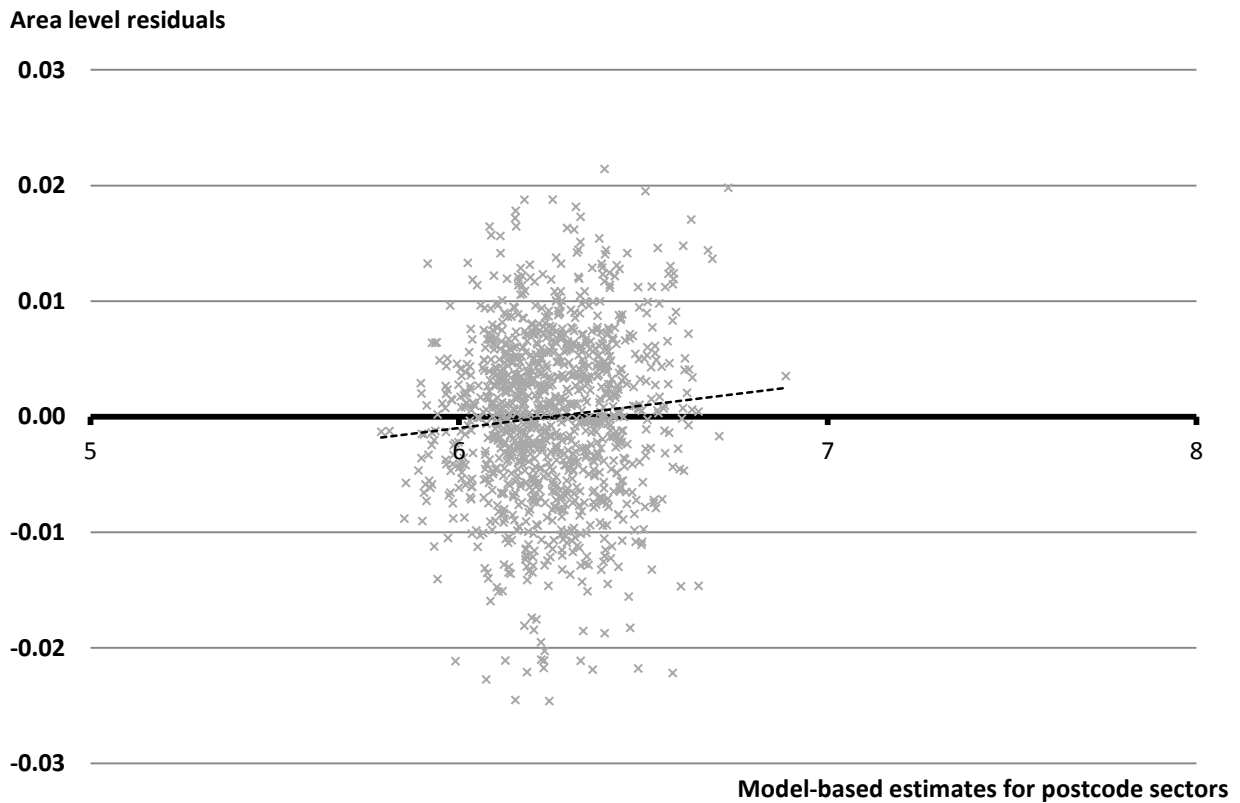


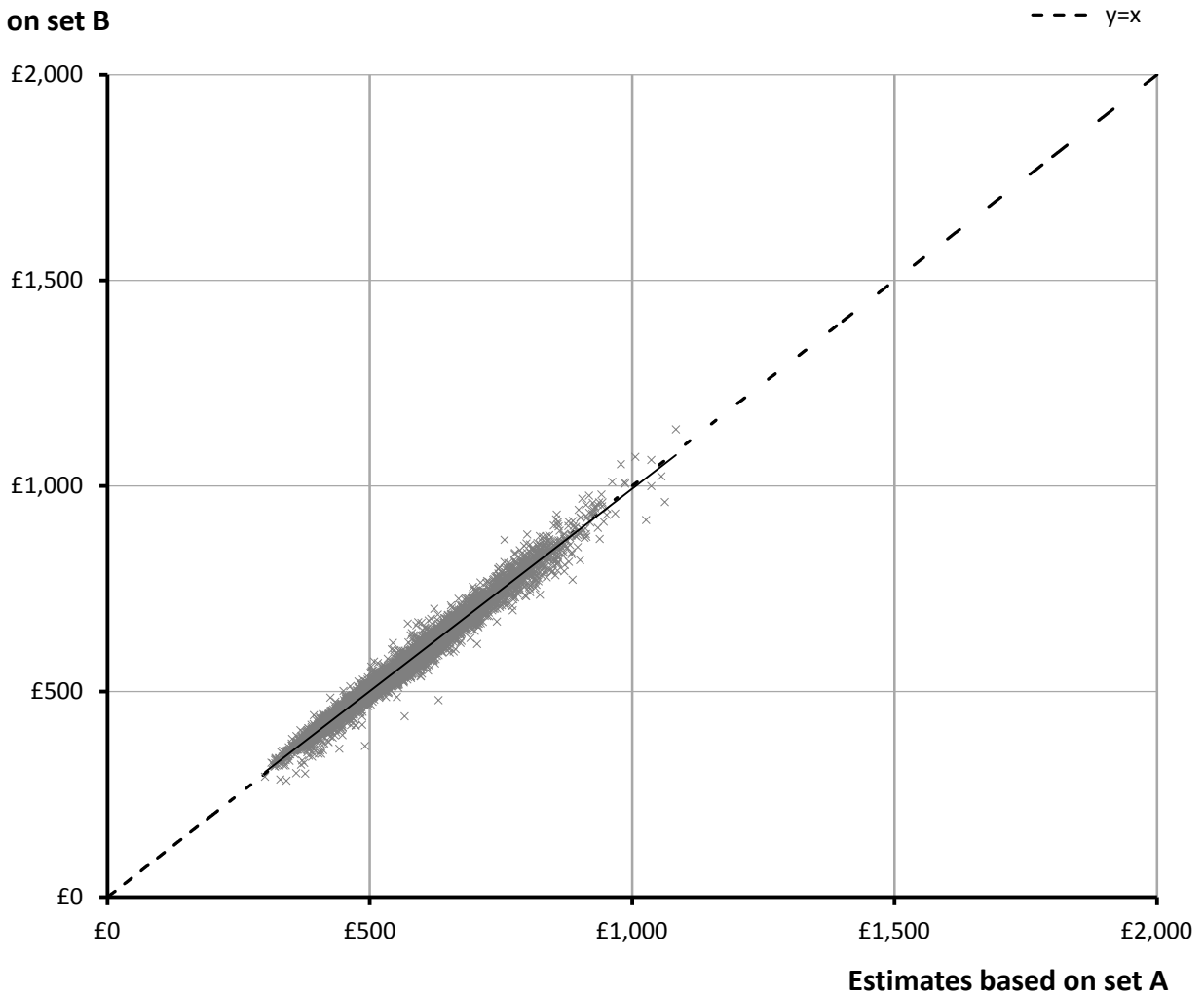
Figure 30: Model-based estimates vs. sample estimates, net income, equivalised before housing costs

Direct survey estimates



Figure 31: Model-based estimates from stability analysis, net income, equivalised before housing costs

Estimates based on set B



Net Weekly Household Income–Equivalised, After Housing Costs

Figure 32: Household level residuals against model-based estimates, net income, equivalised after housing costs

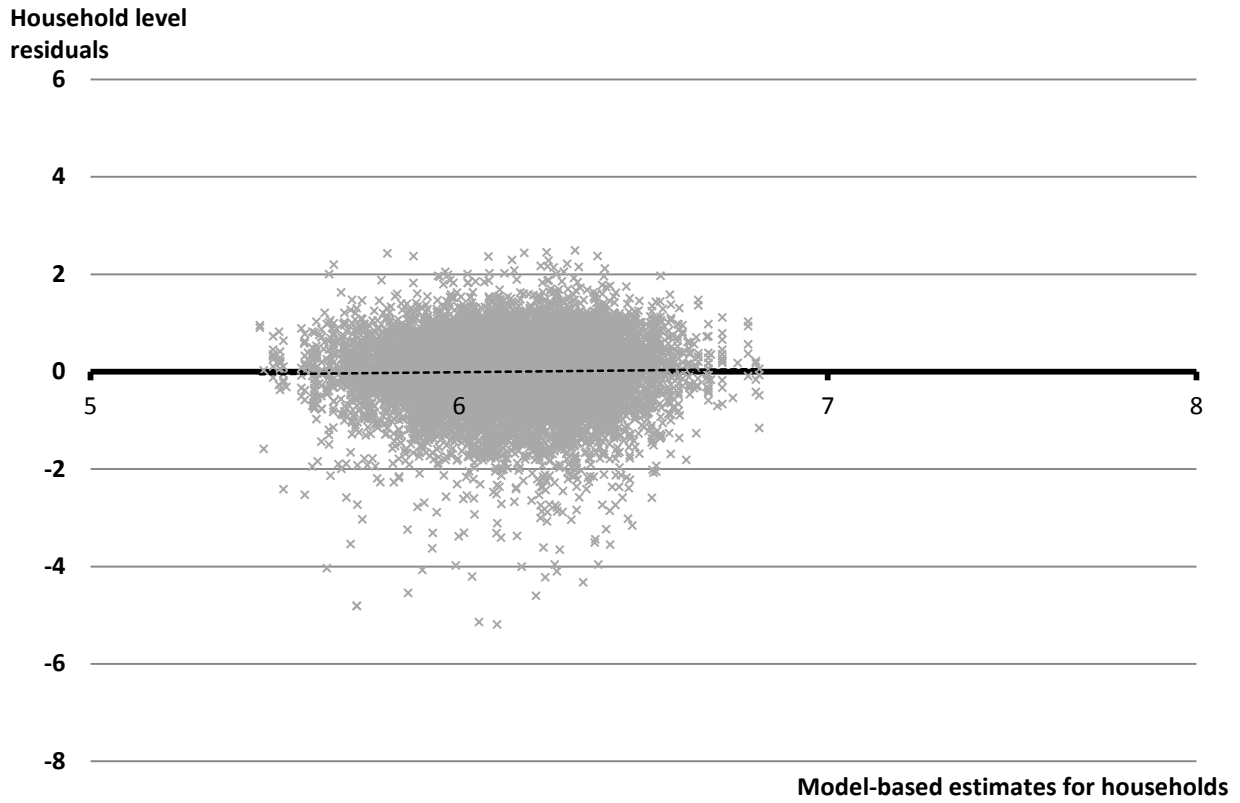


Figure 33: Area level residuals against model-based estimates, net income, equivalised after housing costs

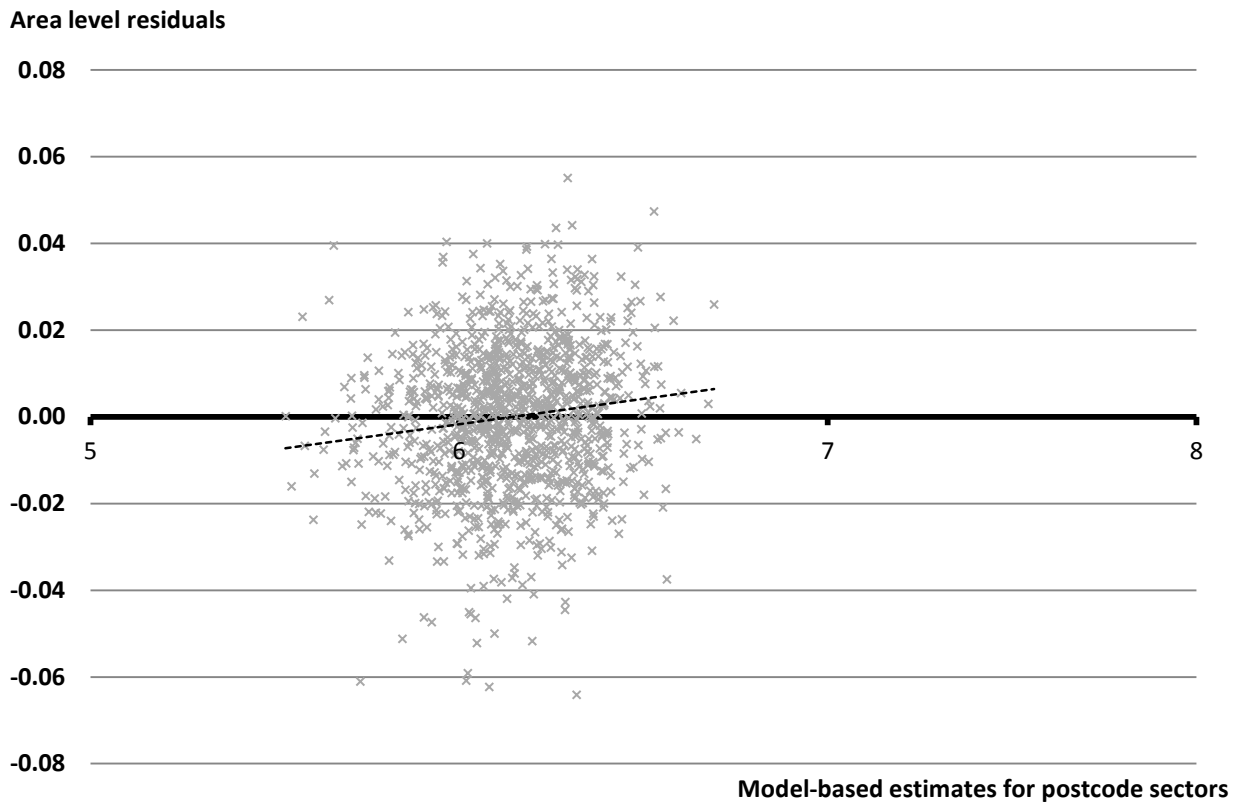


Figure 34: Model-based estimates vs. sample estimates, net income, equivalised after housing costs
Direct survey estimates

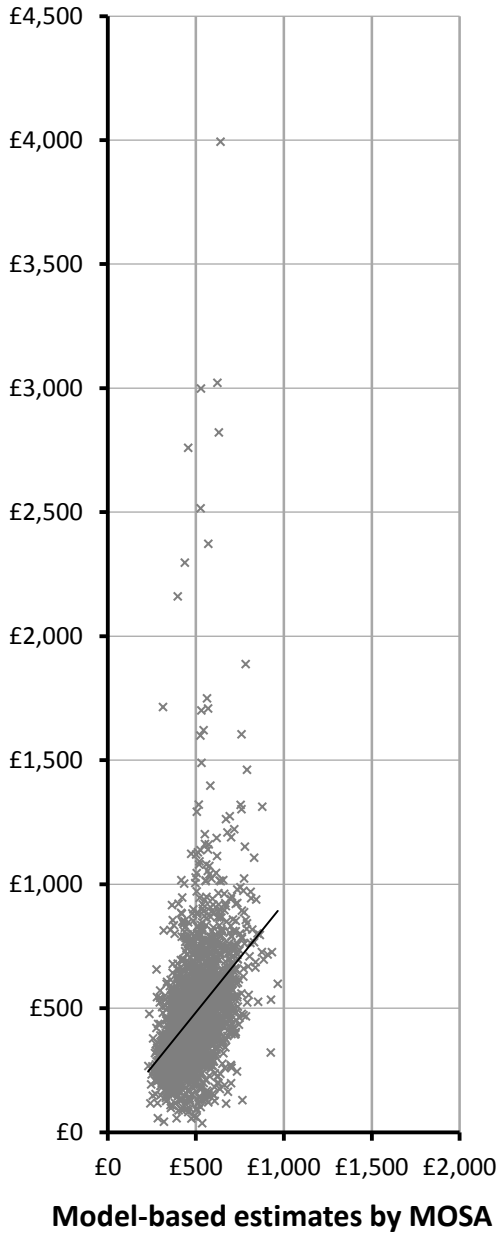
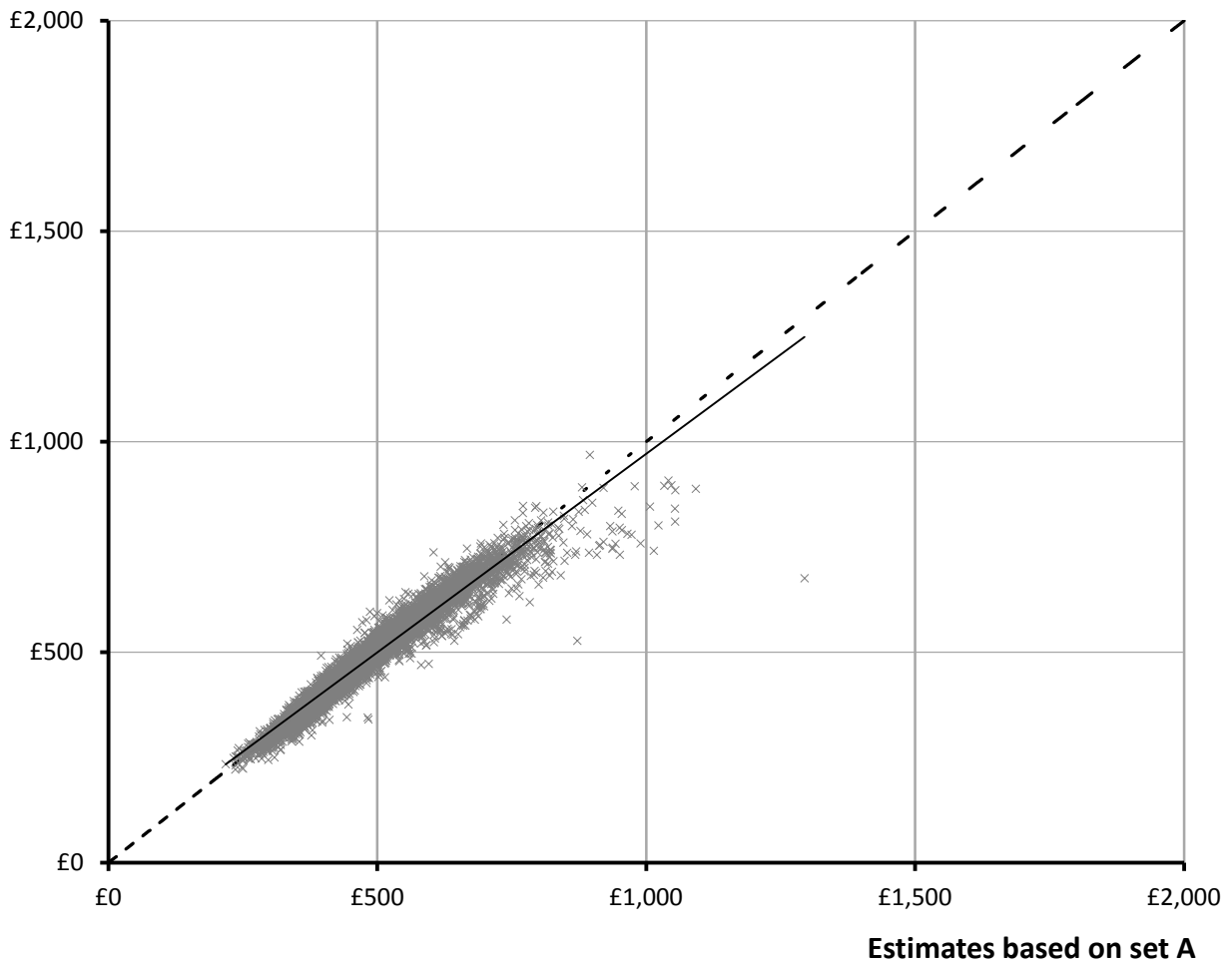


Figure 35: Model-based estimates from stability analysis, net income equivalised, after housing costs

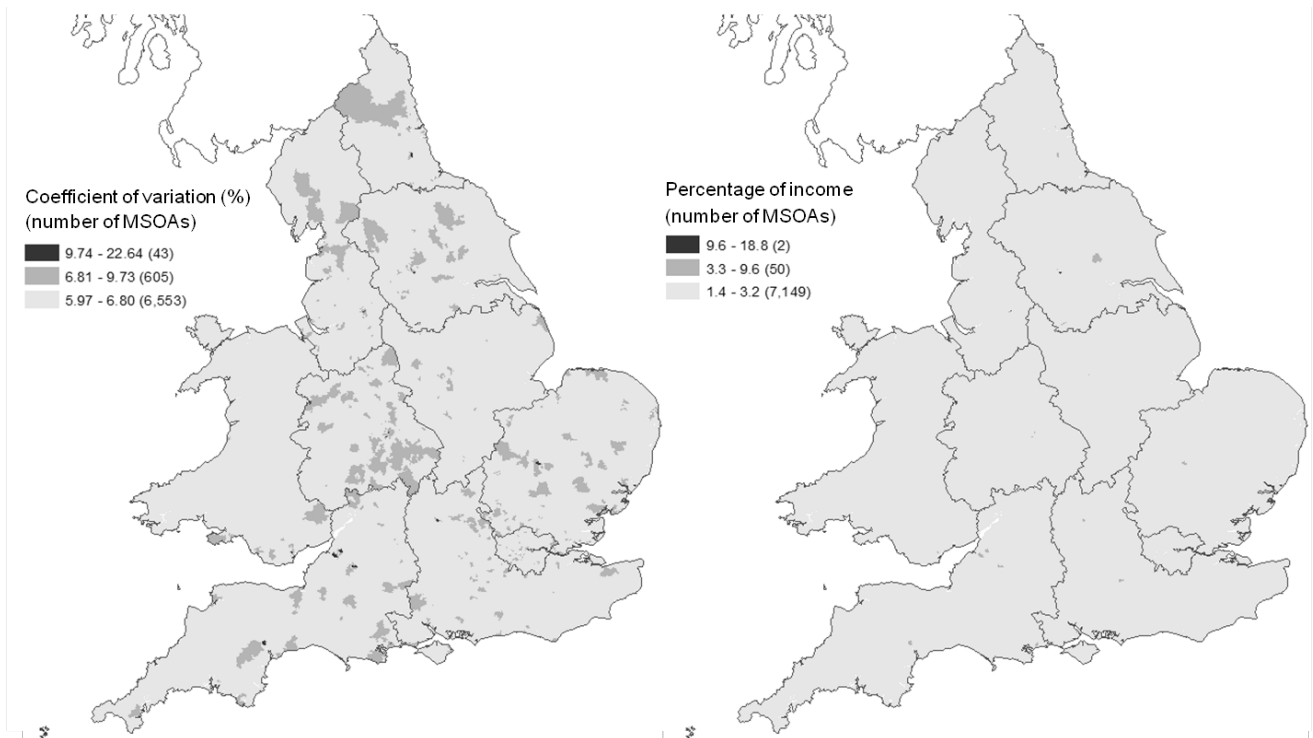
Estimates based on set B

--- $y=x$



Map 7 (left): Coefficient of variation - Net weekly income after housing costs (equivalised) by MSOA, England and Wales, 2013/14;

Map 8 (right): Distance further from estimate of net weekly after before housing costs (equivalised) of the upper confidence limit than the lower confidence limit (expressed as a percentage of the estimate) by MSOA, England and Wales, 2013/14



Source: Office for National Statistics and Ordnance Survey under the Open Government Licence v3.0.
Contains OS data © Crown copyright 2016

F Calculation of Direct Survey Estimates and Confidence Intervals

A number of the diagnostics described in Chapter 6 involve comparing model and direct survey estimates and their standard errors/confidence intervals. This appendix describes the method for calculating the direct MSOA survey estimates and their errors.

When calculating any estimates from the FRS the survey data are grossed up. This is the term given to the process of applying factors to sample data so that they yield estimates for the overall population. The simplest grossing factor system would be a single factor, the uniform grossing factor, which could be calculated as the number of households in the population divided by the number in the achieved sample. However the FRS survey data are grossed by a more complex set of grossing factors, which attempt to correct for differential non-response at the same time as scaling up sample estimates. For more details of this process please refer to the methodology sections of the FRS and HBAI reports (Shale et al (2013) and DWP (2015)). Note that due to the differences in the HBAI and FRS methodology (as described in Appendix B) the two sets of data have different grossing factors.

A model-based approach is adopted to estimate the direct survey estimate for each MSOA and its variance, details of which are provided below. The survey grossing factors are taken into account in these calculations.

The assumption of random area effects implies that the population model appropriate for this situation is where all individuals in the same MSOA (say MSOA k) share a common expected value, $E(y_i) = \mu_k$, a common variance, $Var(y_i) = \sigma_k^2$ and, reflecting the spatial homogeneity of a MSOA, a common covariance, $Cov(y_j, y_j) = \sigma_k^2 \rho_k$. Given a set of sample weights $\{w_i; i \in s\}$ for all individuals in the overall sample s , i.e.

the FRS grossing factors, the approximately model-unbiased direct estimate of the MSOA k mean \bar{y}_k is the weighted mean of all survey responses in that MSOA:

$$\hat{\bar{y}}_k = \hat{N}_k^{-1} \sum_{s_k} w_i y_i = \hat{N}_k^{-1} \hat{t}_k \quad [16]$$

where:

$$\hat{N}_k = \sum_{s_k} w_i \text{ and}$$

s_k denotes the restriction of s to the MSOA k .

Throughout non-informative sampling within a MSOA is assumed, so all inferences relating to y can be conditioned on (the individuals defining) s_k .

To start, it is assumed that the MSOA population size N_k is known and that the MSOA sample size n_k and the sample weights $\{w_i; i \in s\}$ are fixed, so \hat{N}_k is no longer a random variable. Given this setup, it can be shown that the sampling variance of \hat{t}_k is

$$Var(\hat{t}_k - t_k) = Var(\hat{t}_{rk}) - 2Cov(\hat{t}_{rk}, t_{rk}) + Var(t_{rk})$$

where t_k denotes the unknown population total for the MSOA and t_{rk} the corresponding total for the non-sampled individuals in the MSOA (in what follows r_k is used to denote these non-sampled individuals). Also setting $u_i = w_i - 1$, it follows that $\hat{t}_{rk} = \sum_{s_k} u_i y_i$.

Under the above model and conditioning assumptions it follows that:

$$\begin{aligned} Var(\hat{t}_{rk}) &= Var\left(\sum_{s_k} u_i y_i\right) \\ &= \sum_{s_k} u_i^2 Var(y_i) + \sum_{i \in s_k} \sum_{j \neq i \in s_k} u_i u_j Cov(y_i, y_j) \\ &= \sigma_k^2 \left[\sum_{s_k} u_i^2 + \rho_k \sum_{i \in s_k} \sum_{j \neq i \in s_k} u_i u_j \right] \end{aligned}$$

$$\begin{aligned} Var(t_{rk}) &= Var\left(\sum_{r_k} y_i\right) \\ &= \sum_{r_k} Var(y_i) + \sum_{i \in r_k} \sum_{j \neq i \in r_k} Cov(y_i, y_j) \\ &= (N_k - n_k) \sigma_k^2 (1 + \rho_k (N_k - n_k - 1)) \end{aligned}$$

and

$$\begin{aligned} Cov(\hat{t}_{rk}, t_{rk}) &= Cov\left(\sum_{s_k} u_i y_i, \sum_{r_k} y_i\right) \\ &= \rho_k \sigma_k^2 (\hat{N}_k - n_k)(N_k - n_k) \end{aligned}$$

Substituting and collecting terms,

$$Var(\hat{t}_k - t_k) = \sigma_k^2 \left[(1 - \rho_k) \left\{ \sum_{s_k} u_i^2 + N_k - n_k \right\} + \rho_k (\hat{N}_k - N_k)^2 \right] \quad [17]$$

Under the assumed model for the MSOA, a conditionally unbiased estimator of $\sigma_k^2 (1 - \rho_k)$ is

$$\hat{S}_k^2 = (n_k - 1)^{-1} \sum_{s_k} (y_i - \bar{y}_{sk})^2$$

where $\bar{y}_{sk} = n_k^{-1} \sum_{s_k} y_i$ is the unweighted sample mean in MSOA k . That is, with known MSOA population size N_k and assuming that the MSOA specific sample size n_k and sample weights can be considered as fixed, a conditionally unbiased estimator of the sampling variance of \hat{t}_k is

$$\hat{V}(\hat{t}_k) = \hat{S}_k^2 \left[\sum_{s_k} w_i^2 - \hat{N}_k \right] = \hat{S}_k^2 \sum_{s_k} w_i (w_i - 1) \quad [18]$$

Of course, the sample size n_k and the sampling weights are not fixed, but this does not alter the unbiasedness of the above estimator. Since

$$\text{Var}(\hat{t}_k - t_k) = E\left[\text{Var}(\hat{t}_k - t_k \mid \{w_i\}, n_k, N_k)\right] + \text{Var}\left[E(\hat{t}_k - t_k \mid \{w_i\}, n_k, N_k)\right]$$

and $\hat{V}(\hat{t}_k)$ above is an unbiased estimator of the conditional (on n_k and w) sampling variance of \hat{t}_k . It is also an unbiased estimator of the expected value of this conditional variance (the first term on the right hand side above). Furthermore, the corresponding conditional unbiasedness of \hat{t}_k (by construction) means that the second term on the right hand side is zero.

Turning now to the sampling variance of $\hat{y}_k = \hat{N}_k^{-1}\hat{t}_k$ it can be shown that, assuming $E(\hat{N}_k - N_k) = 0$

$$\text{Var}(\hat{y}_k - \bar{y}_k) \approx [E(N_k)]^{-2} \left[\text{Var}(\hat{t}_k - t_k) - 2\mu_k \text{Cov}(\hat{t}_k - t_k, \hat{N}_k - N_k) + \mu_k^2 \text{Var}(\hat{N}_k - N_k) \right]$$

The covariance term can be written as

$$\text{Cov}(\hat{t}_k - t_k, \hat{N}_k - N_k) = E\left[\text{Cov}(\hat{t}_k - t_k, \hat{N}_k - N_k \mid \{w_i\}, n_k, N_k)\right] + \text{Cov}\left[E(\hat{t}_k - t_k \mid \{w_i\}), E(\hat{N}_k - N_k \mid \{w_i\}, n_k, N_k)\right]$$

However, both the terms on the right hand side above are zero, since \hat{N}_k is fixed given the sample weights and \hat{t}_k is conditionally unbiased given these weights. Hence the covariance term is zero.

Turning to the term corresponding to the sampling variance of \hat{N}_k , it is possible to write

$$\hat{N}_k = \sum_s w_i I(i \in k)$$

$$N_k = \sum_U I(i \in k)$$

where $I(i \in k)$ denotes the indicator variable for MSOA k . A working model for this indicator variable is that its values are *iid* Bernoulli with $E[I(i \in k)] = \pi_k$. It immediately follows that

$$\text{Var}(\hat{N}_k - N_k) = \pi_k(1 - \pi_k) \left[\sum_s (w_i - 1)^2 + N - n \right]$$

An unbiased estimator of π_k is $\hat{\pi}_k = N^{-1}\hat{N}_k$ and so an (approximately) unbiased estimator of this sampling variance is

$$\hat{V}(\hat{N}_k) = \hat{\pi}_k(1 - \hat{\pi}_k) \left[\sum_s w_i(w_i - 1) \right]$$

Collecting terms, our final approximately unbiased estimator of the sampling variance of \hat{y}_k is

$$\begin{aligned} \hat{V}(\hat{y}_k) &= \hat{N}_k^{-2} \left[\hat{V}(\hat{t}_k) + \hat{y}_k^2 \hat{V}(\hat{N}_k) \right] \\ &= \hat{N}_k^{-2} \left[\hat{S}_k^2 \sum_s I(i \in k) w_i(w_i - 1) + \hat{y}_k^2 \hat{\pi}_k(1 - \hat{\pi}_k) \sum_s w_i(w_i - 1) \right] \end{aligned} \quad [19]$$

Notes

- Due to the structure of the data an adjustment needs to be made to Equation [19]. The survey sample is based on PCS whereas here we are calculating direct survey estimates for MSOAs which results in small sample sizes, i.e. small n_k , for a majority of MSOAs. Approximately 50% of MSOAs have a sample size of 1 to 5 and small values of n_k will cause \hat{S}_k^2 to be biased/unstable. If $n_k < 5$ then \hat{S}_k^2 is calculated by amalgamating the small sample ward with the nearest (spatially) sampled MSOA.

Equation [19] above requires N to be known in order to calculate π_k . Since here N is unknown it is estimated using the sum of weights across the sample, $\hat{N} = \sum_s w_i$.

G Bibliography

The 2001 Census of Population. (CM 4253). The Stationery Office, 1999. ISBN 0 10 142532 5.

Brown, G., Chambers, R., Heady, P., Heasman, D. (2001).

Evaluation of Small Area Estimation Methods – An Application to Unemployment Estimates from the UK LFS. Proceedings of Statistics Canada Symposium in 2001.

Department for Work and Pensions, FRS (2015).

Family Resources Survey: United Kingdom, 2013/14

John Shale, Khadija Balchin, Juwaria Rahman, Robert Reeve, Mark Rolin (2015)

Households Below Average Income: 1994/95 – 2013/14

Department for Work and Pensions, HBAI team (2013)

Households Below Average Income 1994/95 – 2011/12

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206778/full_hbai13.pdf

Heady, P., Clarke, P., Brown, G., Ellis, K., Heasman, D., Hennell, S., Longhurst, J., Mitchell, B. (2003).

Small Area Estimation Project Report. Model-Based Small Area Estimation Series No.2, ONS Publication.

Longhurst, J., Cruddas, M., Goldring, S., Mitchell, B. (2004).

Model-based Estimates of Income for Wards, 1998/99: Technical Report. Published in Model-Based Small Area Estimation Series, ONS Publication.

Longhurst, J., Cruddas, M., Goldring, S. (2005).

Model-based Estimates of Income for Wards, 2001/02: Technical Report. Published in Model-Based Small Area Estimation Series, ONS Publication.

Office for National Statistics (2005).

Super Output Areas,

<https://www.ons.gov.uk/census/2001censusandearlier/dataandproducts/outputgeography/outputareas>

Teague, A. (1999).

Income Data for Small Area Summary of Response to Consultation. Advisory Group Paper (99)19.

The 2001 Census of Population. (CM 4253). The Stationery Office, 1999. ISBN 0 10 142532 5.