Office for National Statistics

Income estimates for small areas in England and Wales, technical report: financial year ending 2018

Methods used to produce small area income estimates for local areas in England and Wales, known as Middle layer Super Output Areas (MSOAs), including information on the quality of the models and estimates.

Contact:
Sam Cockle
hfig@ons.gov.uk
+44 (0)1633 455779

Release date:
5 March 2020

Next release:
To be announced

Table of contents

1 . Introduction

Importance of income statistics

There is a need for high-quality income statistics at the smallest possible geographical level. Interest in this stems from a variety of sources:

- central government departments
- local authorities
- academics
- commercial organisations
- independent researchers

These data are essential for the identification of deprived and disadvantaged communities, to support work on social exclusion and inequalities, evaluation research, provision of information for practitioners, and the profiling of geographical areas.

Requirement for income data

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 the public acceptability of asking people about their income, and the risks this could have on the overall number of 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.

Use of Middle-layer Super Output Areas

This report is a technical guide to support the financial year ending 2018 (April 2017 to March 2018) set of Middle-layer Super Output Area (MSOA)-level income estimates for England and Wales. 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.

Comparability with other sources

These model-based estimates of average household income in MSOAs are not calculated in the same way as the national and regional household income estimates published separately by the Office for National Statistics (ONS). The definitions of income and data sources used for these statistics are different. It is not possible, therefore, to aggregate the estimates up to match the regional and national 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 that relates the survey variable of interest (for example, 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 the small areas. These estimates and confidence intervals are now published as National Statistics.

Data quality and methods

The report contains details of the methods and processes used, and of the assessment of the quality of the models and the resulting income estimates. Several diagnostic checks are used to assess quality, which 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.

Also included in this report is a comparison of the model (and covariate data) used to derive the income estimates for financial year ending 2018 with that used for financial year ending 2016 and guidance on the use of the estimates.

Further technical information, more background to the need for income estimates for small areas and other methods considered is contained within the previous Technical Report.

2 . Methodology

Synthetic estimation produces estimates for domains where 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 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 (for example, household income) on area-level values of the covariates (for example, proportion of the Middle-layer Super Output Area (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 the Office for National Statistics (ONS) has used in its development of small area estimation. Once a model has been selected an assessment of the quality is made using several diagnostics. For more technical details of the methodology please see the previous Technical Report.

3 . Modelling for income: datasets

Survey data

The survey data were obtained from the 2017 to 2018 Family Resources Survey (FRS). The FRS was chosen as the source for survey data 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 (for example, 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, for example, the estimates produced are values of average Middle-layer Super Output Area (MSOA) income for the following four income types:

- total annual household income (unequivalised)
- net annual household income (unequivalised)
- net annual household income before housing costs (equivalised)
- net annual household income after housing costs (equivalised)

Total annual household income (unequivalised)

This is the sum of the gross income of every member of the household plus any income from benefits, that is, wages and salaries, self-employment, pensions, investments, benefits.

Net annual household income (unequivalised)

This is the sum of the net income of every member of the household, that is, all income minus Income Tax, National Insurance, rates and Council Tax, maintenance and child payments deducted through pay, contribution to students living away, contributions to occupational pensions.

Net annual household income before housing costs (equivalised)

This is the same as net annual household income unequivalised but is then subject to an equalisation scale.

Net annual household income after housing costs (equivalised)

This uses the same elements as net annual household income but also deducts housing costs, that is, rent, water rates, mortgage interest payments, structural insurance premiums, ground rent and service charges prior to the equivalisation scale.

Equivalisation

Equivalised 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. Equivalisation is needed to make sensible income comparisons between households. These estimates use the Organisation for Economic Co-operation and Development (OECD) equivalisation scale, as is standard across other Office for National Statistics (ONS) income measures. For more details on these income definitions and the equivalisation scale see the previous Technical Report.

As was the case in financial year ending 2016, we have published the financial year ending 2018 incomes estimates in terms of annual income (rounded to the nearest £100) rather than weekly income (rounded to the nearest £1) to aid interpretation. However, the estimates were modelled using weekly income data as per previous outputs. The final weekly estimates are expressed as annual income using a factor of 365.24/7.

Sample size

The FRS uses a stratified clustered probability sample drawn from the Royal Mail's 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 financial year ending 2018, 26 addresses were selected per PSU (for April to December, with 28 addresses per PSU for January to March). More information on the FRS methodology is contained within the FRS Background note and methodology report.

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 (for example, excluding the assets section of the questionnaire). In financial year ending 2018 the achieved sample size (for the UK) was 19,136 households.

Survey data file

The requirement for this release is to produce MSOA-level estimates of average household income (four types) for England and Wales. The survey data file used contained 14,471 households from 1,171 postcode sectors in financial year ending 2018. The final survey data file for England and Wales contained cases in 2,555 different MSOAs out of a total of 7,201. The number of cases per MSOA in the achieved FRS sample varies widely particularly because MSOAs cut across the postcode sectors' primary sampling unit. For example, some MSOAs recorded only one response whereas others had 34 (the maximum number of sampled households).

For each different income type a few records were found with values of income less than or equal to £1, these were removed from the sample dataset. Additional records with extremely high total income values were removed as they would have had an unduly large influence on the model [1]. For the net weekly (unequivalised and equivalised) income, records were removed where the net income was greater than the total income by £10. The net equivalised 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 data (HBAI) [2].

Definitions from FRS data

Although all the survey data used in the modelling process are obtained from the FRS, two of these income types are defined by a different study that is based on FRS data. Net weekly household income (equivalised) both before and after housing costs is defined and calculated in the HBAI report.

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 dataset. The HBAI dataset 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. Note that because of the differences in the HBAI and FRS methodology the two sets of data have different grossing factors.

Covariate datasets

The 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 2017 (provided as counts; see subheading DWP data)
- Valuation Office Agency (VOA) Council Tax Bandings, March 2017 (provided as counts and transformed into proportions; see subheading VOA Council Tax bandings)
- Office for National Statistics, House Price Statistics for Small Areas, Quarter 1 (January to March) 2018 (in addition to counts of the number of dwelling sales, data contain measures of house prices (median price) for sales that took place)
- Department of Energy and Climate Change, Energy Consumption data, 2017
- Her Majesty's Revenue and Customs, Pay as You Earn data, 2018
- Regional or country identification variable (for more information see subheading Regional or country identification variable)

DWP data

The DWP data were provided as counts. However, it was more appropriate to include proportions or prevalence rates in the modelling process. MSOA population data from mid-2017 were used as denominators to derive these proportions.

VOA Council Tax bandings

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.

Regional or 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 and country, and 0 otherwise. These region and country variables are listed in Table 1. Note that London was selected as the base case and therefore not specified separately in the modelling procedure.

Table 1: Regional variables included in modelling income

Variable name	Country or region
northest	North East
northwst	North West
york	Yorkshire and The Humber
eastmid	East Midlands
westmid	West Midlands
east	East of England
southest	South East
southwst	South West
wales	Wales

Source: Office for National Statistics

The data used are as close to the reference period of the target income estimates as possible (that is, for financial year ending 2018). Administrative data are collected primarily for government administrative processes and may change over time.

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 dataset comprises the survey response variables of interest, weekly household income, matched to postcodes, and MSOA codes, for the estimation area. The covariate dataset comprises MSOA covariates along with the corresponding MSOA identifiers. These two datasets are matched by reference to the MSOA codes.

The resulting matched dataset, containing the survey variable along with associated covariates and MSOA and PCS identifiers, becomes the analysis dataset. The analysis dataset is required for the modelling and the full covariate dataset is required to produce the final estimates once the modelling has been performed.

Notes for: Modelling for income: datasets

1. These households either had a total weekly household income that 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 MSOA.

2. The Households below average income dataset is a record level dataset maintained by the Department for Work and Pensions.

4 . 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.

The developed models are fitted as multilevel models and can be used to produce estimates of the target variable at the small area level, for example, the models can be used to produce Middle-layer Super Output Area (MSOA)-level estimates of average weekly household 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 and country indicator terms are forced into the model (whether significant or not) and then the method of step-wise forward selection is used to identify the significant covariates to be included in the models from the set of covariates.

All 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, for example, 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. Note covariates are sometimes 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 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 Family Resources Survey (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 and country level, and comparisons made between the two sets of estimates. The ratio of direct survey estimate to aggregated model estimate at the region and 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 the previous Technical Report.

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

Total weekly household income (unequivalised)

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

Covariate Name	Label	Source	T ratio
northest	North East	Country/regional indicators	-1.24
northwst	North West	Country/regional indicators	-2.05
york	Yorkshire and The Humber	Country/regional indicators	-2.04
eastmid	East Midlands	Country/regional indicators	-1.14
westmid	West Midlands	Country/regional indicators	-1.8
east	East of England	Country/regional indicators	-0.83
southest	South East	Country/regional indicators	-0.96
southwst	South West	Country/regional indicators	-1.75
wales	Wales	Country/regional indicators	-0.97
lnpctg_s	Logit of proportion of people claiming Pension Credit, Award type – guarantee element and saving element	DWP	-4.06
ewlnpayet2	Log of the median annual gross individual PAYE income	HMRC	3.55
lnpgroupab	Logit of proportion of people aged 16 to 74 whose approximated social grade is AB	Census	6.27
phhtype1	Proportion of households that contain one person only	Census	-2.2
ewdlaa1	Proportion of people claiming Disability Living Allowance aged under 25	DWP	2.36
ewpayeg5tp	Tenth Percentile of annual gross individual PAYE income – Females, aged 60-64	HMRC	2.77
wdef_l	Proportion/count of dwellings in Welsh Council Tax bands D, E and F	VOA Council Tax Data	2.22
pcommun	Proportion of people living in communal establishments	Census	-2.2
lnphrpreli	Logit of proportion of household reference persons who have a religion	Census	-1.21
lnpgroupab_eastmid	Interaction between lnpgroupab (Logit of proportion of people aged 16 to 74 whose approximated social grade is AB) and eastmid (East Midlands)	Census & Country /regional indicators	2.91
lnphrpreli_wales	Interaction between lnphrpreli (Logit of proportion of household reference persons who have a religion) and Wales	Census & Country /regional indicators	-2.49
pcommun_wales	Interaction between pcommun (Proportion of people living in communal establishments) and Wales	Census & Country /regional indicators	2.28
ewlnpayet2_southest	Interaction between ewinpayet2 (Log of the median annual gross individual PAYE income) and southest (South East)	Admin Data & Country/regional indicators	2.34

Source: Office for National Statistics

With no covariates included in the model the estimated standard residual area variance $\hat{\sigma}_u^2$

is 0.0418 (0.0038) compared with 0.0075 (0.0023) when the significant covariates are included in the model, a decrease of 82.12%. Therefore, these covariates together account for 82.12% of the total between area variance.

The most significant covariate in the model is the census covariate “lnpgroupab”, which has a T value of 6.27. As one would expect this covariate has a positive coefficient; as the proportion of people aged 16 to 74 years whose approximated social grade is AB increases so does the average household income for that MSOA.

“ewlnpayet2” is the next most significant covariate in the model with a positive coefficient, this has a T-value of 3.55. It shows that as the median annual gross individual Pay As You Earn (PAYE) income amount increases so does the average household income.

The relationship of a covariate with the average household income may be different if it is also involved in a model interaction. For example, “pcommun” is included in a model interaction with “wales”. This suggests that the relationship between “pcommun” and the average household income is different for Wales MSOAs.

The most significant variable with a negative coefficient was “lnpctg_s”, which refers to the logit of proportion for the Pension Credit Award type – guarantee element and saving element, which has a T value of negative 4.06. This means that a lower proportion of people claiming this benefit within an area is associated with a higher average household income.

Net weekly household income (unequivalised)

Table 3: Key to covariates included in the model for net weekly household income (unequivalised)

Covariate Name	Label	Source	T ratio
northest	North East	Country /regional indicators	-1.6
northwst	North West	Country /regional indicators	-1.88
york	Yorkshire and The Humber	Country /regional indicators	-1.35
eastmid	East Midlands	Country /regional indicators	-2.82
westmid	West Midlands	Country /regional indicators	-1.3
east	East of England	Country /regional indicators	-0.79
southest	South East	Country /regional indicators	-0.42
southwst	South West	Country /regional indicators	-1.42
wales	Wales	Country /regional indicators	-1.47
lnpctg_s	Logit of proportion of Pension Credit Award type – guarantee element and saving element	DWP	-3.58
ewlnpayet2	Log of the median annual gross individual PAYE income	HMRC	3.59
ewdlaa1	Proportion of people claiming Disability Living Allowance - Aged under 25	DWP	2.51
lnpcp	Logit of proportion of Pension Credit – Partner	DWP	3
lnpcd3	Logit of proportion of Pension Credit - Claim duration 2-5 years	DWP	-2.96
lnibsdtot	Logit of proportion of Incapacity Benefit / Severe Disablement Allowance claimants – Total	DWP	3.7
ewpayeg1md	Median annual gross individual PAYE income – Male, aged 16-59	HMRC	-2.39
lnphrpreli	Logit of proportion of household reference persons who have a religion	Census	-2.24
wabc_l	Proportion/count of dwellings in Welsh Council Tax bands A, B and C	VOA Council Tax Data	-2.43
pcommun	Proportion of people living in communal establishments	Census	-4.3
pgroupab	Proportion of people aged 16 to 74 whose approximated social grade is AB	Census	5.76
lnpgroupc2	Logit of proportion of people aged 16 to 74 whose approximated social grade is C2	Census	4.5

ewaelecc	Average Consumption of Ordinary Domestic Electricity	Department of Energy & Climate Change	-2.03
ewlnpayeg4lq	Log of lower quartile annual gross individual PAYE income – Female, aged 16-59 for the 2017/18 tax year	HMRC	-1.81
lnpchbath	Logit of proportion of households with sole use of a bath/shower and toilet and central heating	Census	2.45
lncatot	Logit of proportion of Carers Allowance claimants - Total	DWP	-3.08
lndlaa5	Logit of proportion of Disability Living Allowance claimants – Aged 70+	DWP	2.13
lnpcd3_eastmid	Interaction between lnpcd3 (Logit of proportion of Pension Credit – Partner) and eastmid (East Midlands)	DWP & Country /regional indicators	-3.3
pcommun_wales	Interaction between pcommun (Proportion of people living in communal establishments) and Wales	DWP & Country /regional indicators	2.81
ewlnpayeg4lq_lnibsdtot	Interaction between ewlnpayeg4lq (Log of lower quartile of annual gross individual PAYE income – Female, aged 16-59 and lnibsdtot (Logit of proportion of Incapacity Benefit / Severe Disablement Allowance claimants – Total	HMRC & DWP	-2.63
lnphrpreli_wales	Interaction between lnphrpreli (Logit of proportion of household reference persons who have a religion) and Wales	DWP & Country /regional indicators	-2.13
lnibsdtot_westmid	Interaction between lnibsdtot (Logit of proportion of Incapacity Benefit / Severe Disablement Allowance claimants – Total) and westmid (West Midlands)	DWP & Country /regional indicators	-2.33
ewpayeg1md_york	Interaction between ewpayeg1md (Median annual gross individual PAYE income amount by MSOA – Male, aged 16-59) and york (Yorkshire and The Humber)	HMRC & Country /regional indicators	3.31
lnpcd3_york	Interaction between lnpcd3 (Logit of proportion of Pension Credit – Partner) and york (Yorkshire and The Humber)	DWP & Country /regional indicators	3.07
pcommun_york	Interaction between pcommun (Proportion of people living in communal establishments) and Yorkshire and The Humber	DWP & Country /regional indicators	2.58

Source: Office for National Statistics

With no covariates included in the model the estimated residual area variance
$\hat{\sigma}_u^2$

is 0.0299 (0.003) compared with 0.0058 (0.002) when the significant covariates are included in the model, a decrease of 80.54%. Therefore, these covariates together accounted for 80.54% of the total between area variance.

The most significant covariate in the model is the census covariate “pgroupab”, (proportion of people aged 16 to 74 years whose approximated social grade is AB), which has a T value of 5.76. As one would expect this covariate has a positive coefficient; as the proportion of people aged 16 to 74 years whose approximated social grade is AB increases so does the average household income for that MSOA.

“lnpgroupc2” is the next most significant covariate in the model with a positive coefficient and has a T-value of 4.50. This also shows that, as the proportion of people aged 16 to 74 years whose approximated social grade is C2 in an MSOA increases, the average household income increases.

The most significant variable with a negative coefficient was “pcommun”, which has a T value of negative 4.30. This shows that as the proportion of people living in communal establishments increases, the average household income for that MSOA decreases.

Net weekly household income – equivalised, before housing costs

Table 4: Key to covariates included in the model for net weekly household income, equivalised, before housing costs

Covariate Name	Label	Source	T ratio
northest	North East	Country /regional indicators	-2.71
northwst	North West	Country /regional indicators	-1.53
york	Yorkshire and The Humber	Country /regional indicators	-2.26
eastmid	East Midlands	Country /regional indicators	-2.23
westmid	West Midlands	Country /regional indicators	-1.38
east	East of England	Country /regional indicators	-0.65
southest	South East	Country /regional indicators	-0.26
southwst	South West	Country /regional indicators	-0.9
wales	Wales	Country /regional indicators	-1.78
lnpcd3	Logit of proportion of Pension Credit - Claim duration 2-5 years	DWP	-2.01
page16un	Proportion of people aged under 16	Census	-3.76
ewlnpayeg3md	Log of the median of annual gross individual PAYE income amount by MSOA – Male, aged 65+	HMRC	4.83
ewpcts	Proportion of people claiming Pension Credit – Award type Saving Element Only	Census	3.24
lnpgroupc2	Logit of proportion of people aged 16 to 74 whose approximated social grade is C2	Census	3.57
ewspf	Proportion of people claiming State Pension - Female	DWP	-5.01
lnpchbath	Logit of proportion of households with sole use of a bath/shower and toilet and central heating	Census	2.86
ewpayeg4tp	Tenth percentile of annual gross individual PAYE income amount by MSOA – Female, aged 16-59 for the 2017/18 tax year	HMRC	1.96
pcommun	Proportion of people living in communal establishments	Census	-3.08
pgroupab	Proportion of people aged 16 to 74 whose approximated social grade is AB	Census	6.75
wdef_l	Proportion/count of dwellings in Welsh Council Tax bands D, E and F	VOA Council Tax Data	2.38
ewavhhroom	Average number of rooms per household	Census	-2.45
ewspa2	Proportion of people claiming State Pension – Aged 70+	DWP	0.83

ewspa2_eastmid	Interaction between ewspa2 (Proportion of people claiming State Pension – Aged 70+) and eastmid (East Midlands)	DWP & Country /regional indicators	3.08
lnpcd3_eastmid	Interaction between lnpcd3 (Logit of proportion of Pension Credit – Partner) and eastmid (East Midlands)	DWP & Country /regional indicators	-2.72
ewspf_york	Interaction between ewspf (Proportion of people claiming State Pension - Female) and york (Yorkshire and The Humber)	DWP & Country /regional indicators	2.37
ewpayeg4tp_southest	Interaction between ewpayeg4tp (Tenth percentile of annual gross individual PAYE income amount by MSOA Female, aged 16-59 for the 2017/18) and southest (South East)	HMRC & Country /regional indicators	3.07
lnpchbath_southest	Interaction between lnpchbath (proportion of households with sole use of a bath/shower and toilet and central heating) and southest (South East)	Census	-2

Source: Office for National Statistics

With no covariates included in the model the estimated residual area variance
$\hat{\sigma}_u^2$

was 0.0239 (0.002) compared with 0.004 (0.001) when the significant covariates were included in the model, a decrease of 85.30%. Therefore, these covariates together accounted for 85.30% of the total between area variance.

The most significant covariate in the model is the census covariate, “pgroupab” (proportion of people aged 16 to 74 years whose approximate social grade is AB), which has a T value of 6.75. As one would expect this covariate has a positive coefficient; as the proportion of people of this social grade increases so does the average household income for that MSOA.

“ewlnpayeg3md” is the next most significant covariate in the model with a positive coefficient and has a T value of 4.83. This shows that, as the log of the median annual gross individual PAYE income for males aged over 65 years in an MSOA increases, the average household income increases.

The most significant variable with a negative coefficient was “ewspf” (proportion of females claiming State Pension), which has a T value of negative 5.01. This shows that as the proportion of females claiming State Pension increases, the average household income for that MSOA decreases.

Net weekly household income – equivalised, after housing costs

Table 5: Key to covariates included in the model for equivalised net weekly household income after housing costs

Covariate Name	Label	Source	T ratio
northest	North East	Country /regional indicators	-2.33
northwst	North West	Country /regional indicators	-0.85
york	Yorkshire and The Humber	Country /regional indicators	-1.34
eastmid	East Midlands	Country /regional indicators	-2.19
westmid	West Midlands	Country /regional indicators	-1.17
east	East of England	Country /regional indicators	-2.01
southest	South East	Country /regional indicators	-1.3
southwst	South West	Country /regional indicators	-1.84
wales	Wales	Country /regional indicators	-1.89
ewlnpayeg3md	Log of the median annual gross individual PAYE income – Males, aged 65+	HMRC	4.77
page16un	Proportion of people aged under 16	Census	-4.85
lnpcd3	Logit of proportion of Pension Credit - Claim duration 2-5 years	DWP	-3.07
pcommun	Proportion of people living in communal establishments	Census	-4.37
ewpcts	Proportion of people claiming Pension Credit – Award type saving element only	DWP	3.07
lnpgroupab	Logit of proportion of people aged 16 to 74 whose approximated social grade is AB	Census	5.83
lnphhtype6	Logit of proportion of households that are a couple with dependent child (ren)	Census	3.84
lnengabc	Logit of proportion of dwellings in England, council tax bands A, B and C	VOA Council Tax Data	4.2
ewdlaa1	Proportion of people claiming Disability Living Allowance aged under 25	DWP	2.82
lncam	Logit of proportion of Carers' Allowance claimants - Male	DWP	-3.87
lnpgroupc2	Logit of proportion of people aged 16 to 74 whose approximated social grade is C2	Census	2.29
lnpchbath	Logit of proportion of households with sole use of a bath/shower and toilet and central heating	Census	2.06

lnpcd3_eastmid	Interaction between lnpcd3 (Logit of proportion of Pension Credit – Partner) and eastmid (East Midlands)	DWP & Country /regional indicators	-4.03
lnpchbath_lncam	Interaction between lnpchbath (Logit of proportion of households with sole use of a bath/shower and toilet and central heating) and lncam (Logit of proportion of Carers Allowance claimants – Male)	Census & DWP	3.24
lnpchbath_lnpcd3	Interaction between lnpchbath (Logit of proportion of households with sole use of a bath/shower and toilet and central heating) and lnpcd3 (Logit of proportion of Pension Credit – Partner)	Census & DWP	-2.27
pcommun_wales	Interaction between pcommun (Proportion of people living in communal establishments) and Wales	Census & Country /regional indicators	2.02

Source: Office for National Statistics

With no covariates included in the model the estimated residual area variance $\hat{\sigma}_u^2$

was 0.029 (0.003) compared with 0.001 (0.002) when the significant covariates were included in the model, a decrease of 96.3%. Therefore, these covariates together accounted for 96.3% of the total between area variance.

The most significant covariate in the model is the census covariate "'lnpgroupab", which has a T value of 5.83. As the proportion of people aged 16 to 74 years whose approximated social grade is AB increases so does the average household income for that MSOA.

"ewlnpayeg3md" is the next most significant covariate in the model with a positive coefficient and has a T value of 4.77. This shows that, as the log of the median annual gross individual PAYE income for males aged over 65 years increases, the average household income increases.

The most significant variable with a negative coefficient was "page16un", which has a T value of negative 4.85. This shows that as the proportion of people aged under 16 years increases, the average household income for that MSOA decreases.

Observations

As expected, 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 of the models the most significant covariate is one of the census covariates, the proportion of people whose approximate social grade is AB. This covariate has a positive coefficient, meaning as the proportion of people of this social grade increases so does the average weekly household income for that MSOA.

2. The majority of regional or 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 or 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.

5 . Quality of the estimates

Once a model has been selected an assessment of the quality is made using several diagnostics. Different diagnostic checks have been used to assess the appropriateness of the models developed. The diagnostic checks employed here are those developed by the Office for National Statistics (ONS) for small area estimation as well as some additional ones. The analysis shows that in general the models are well specified and the assumptions are satisfied. This provides confidence in the accuracy of the estimates and their confidence intervals produced from the models.

The results of the diagnostics for all four income types are summarised in Table 6. Some further information about each test is given below the table. More detail about the diagnostic tests and why they are performed can be found in the previous Technical Report.

Table 6: Diagnostic results for all four income types estimated, England and Wales

Diagnostic Measure		Total Weekly Household Income (unequivalised)	Net Weekly Household Income (unequivalised)	Net Weekly Household ncome, Equivalised Before Housing costs	Net Weekly Household Income, Equivalised After Housing costs
Residual vs Model Estimates	**Constant (SE)**	-0.992 (0.204)	-1.036 (0.210)	-0.782 (0.195)	-0.191 (0.193)
Household Level Residuals	**Slope (SE)**	0.149 (0.030)	0.162 (0.033)	0.123 (0.031)	0.031 (0.031)
Residual vs Model Estimates	**Constant (SE)**	-0.189 (0.032)	-0.188 (0.031)	-0.119 (0.024)	-0.010 (0.006)
Area Level Residuals	**Slope (SE)**	0.028 (0.005)	0.029 (0.005)	0.019 (0.004)	0.002 (0.001)
Model vs Sample Estimates	**Constant (SE)**	-118.058 (40.880)	-24.607 (30.970)	-24.706 (28.037)	10.606 (25.986)
	Slope (SE)	1.122 (0.048)	0.973 (0.046)	0.993 (0.045)	0.948 (0.047)
Model vs Sample Estimates	**Constant (SE)**	240.170 (145.318)	160.23 (129.015)	55.210 (111.841)	42.801 (96.162)
	Slope (SE)	0.277 (0.332)	0.426 (0.374)	0.736 (0.351)	0.827 (0.351)
	Quadratic term (SE)	0.0005 (0.0002)	0.0004 (0.0003)	0.0002 (0.0003)	0.0001 (0.0003)
Coverage	**%**	100	100	99.96	99.92
Wald	**P-value**	1	1	1	1
Stability Analysis	**RRMSE**	0.05	0.066	0.049	0.052

Source: Office for National Statistics

The following paragraphs describe some of the diagnostic tests performed on the data. If you require more detail, please see the previous Technical Report or contact us.

Residual compared with 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 misspecification and non-constant variance of the residuals (heteroscedasticity). If any pattern remains in the residuals this implies model misspecification, for example, 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. The standard errors can be used to determine whether the constant and linear terms are significantly different from 0.

Model compared with 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 misestimated, 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 the previous Technical Report, along with further detail about this diagnostic test. The results show that in quadratic fit, the quadratic term is not significant and neither is the intercept. In the linear fit, the intercept term is not significantly different from 0 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.

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 the sample, there will be direct survey estimates with associated 95% confidence intervals. The diagnostic measures the overlap between the direct confidence intervals and the corresponding model-based estimate confidence intervals, for example, 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 (that is, narrow the width) of the confidence intervals being compared to ensure a 95% overlap. Further details of the modification and this test are available in the previous Technical Report. Any significant deviation from a 95% overlap indicates that the model-based confidence intervals are generally too wide or too narrow.

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 caused by the variances for the direct estimates.

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 and there should be a non-significant p-value associated with the Wald statistic. 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.

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, 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 to obtain regression coefficients.

In a similar way, the other half of the data is used in the model to obtain the regression coefficients. 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) 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. An RRMSE of greater than 0.5 is considered here as an indication of instability. The stability analyses for the four models indicate that the different sets of data produce similar sets of estimates.

6 . Comparing results for financial years ending 2016, 2018 and measuring change

Middle-layer Super Output Area (MSOA)-level model-based estimates of average annual household income have been produced for financial year 2018 in England and Wales, fulfilling users' requirements for income information at MSOA level.

Models

In financial year ending 2016, each model related the Family Resource Survey (FRS) survey estimate of weekly household income to the following covariate:

- Region and country in which MSOA lies

In financial year ending 2018, each model contained the following covariates:

- Region and country in which MSOA lies
- Logit of proportion and proportion of people aged 16 to 74 years whose approximated social grade is AB
- Proportion of people living in communal establishments

Diagnostics

Some plots of household- and area-level residuals for all models, for both financial years ending 2016 and 2018, 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.

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 caused by overestimating 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 financial years ending 2016 and 2018 models produce fairly consistent results indicating that in general the models for England and Wales are well-specified and the assumptions are satisfied, although not as strongly in financial year ending 2016 as previous years, or as in financial year 2018. This produces confidence in the accuracy of the estimates and their confidence intervals produced from the models.

Estimates

The different models described previously 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.

Covariates

The following covariates were available for modelling the financial years ending 2016 and 2018 MSOA model-based estimates of income.

Table 7: Covariate data used in the models for the financial years ending 2016 and 2018 model-based estimates of income

Financial year ending 2016	Financial year ending 2018
Census data, 2011	Census data, 2011
HMRC tax credits data, 2015; HMRC PAYE and SHBE data, March 2016	HMRC PAYE data, March 2018
DWP benefit data, 2015	DWP benefit data, 2017
Region/country indicators	Region/country indicators
ONS House Price Statistics for Small Areas year ending March 2016	ONS House Price Statistics for Small Areas year ending March 2017
Council Tax data, March 2015	Council Tax data, March 2017
DECC Energy Consumption data 2015	DECC Energy Consumption data 2017

Source: Office for National Statistics

Table 7 shows that different covariate data-sets were available and used at the time of modelling the financial years ending 2016 and 2018 model-based estimates of average income.

Different covariates have been selected in the models for financial years ending 2016 and 2018. This is both a consequence of the covariate selection process as well as the availability of different covariate datasets 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 financial years ending 2016 and 2018 could partly be because of differences in the covariates selected in the models rather than a real change in the mean household income for that area.

Geography of estimation

The financial year ending 2002 income estimates were produced on 2003 Census Area Statistics (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), for example, the majority of CAS wards are physically different to MSOAs. Comparisons between financial year ending 2002 estimates and later estimates are therefore not usually possible because of boundary differences.

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 financial years ending 2005, 2008, 2012, 2014, 2016 and 2018 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 two different time periods.

However, the financial years ending 2012, 2014, 2016 and 2018 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 these estimates and earlier estimates is not possible. In these instances, the geography code, which represents the MSOA for financial year ending 2018 (or 2016 or 2014 or 2012), 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.

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

Decrease in precision of estimates for financial year ending 2016

The models for financial year ending 2016 were found to account for less of the between area variance compared with previous years (further detail is available in the previous Technical Report). This was not the case for the current financial year ending 2018 models, which were more aligned to previous years.

7 . Guidance on the use of the estimates

The results of the diagnostic checks presented previously 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 most cases, the estimates are produced using the most up-to-date covariate data sources to match the financial year ending 2018 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.

Estimates can be used to make inferences such as the average 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 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, for example, a MSOA may have a larger modelled estimate for net weekly household income (unequivalised) when compared with total household income (unequivalised). Although there may be some such inconsistencies the models selected are the best possible to model the general patterns of income overall 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 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 financial year ending 2018 survey data and so are only valid for this period.