Income estimates for small areas in England and Wales, technical report: financial year ending 2020

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

Contact:
Andrew Zelin
Economic.Wellbeing@ons.gov.uk
+44 1329 447767

Release date:
11 October 2023

Next release:
To be announced

Table of contents

1 . Overview of income estimates

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

Questions on income have never been included in the UK 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 (FYE) 2020 (April 2019 to March 2020) 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 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 number of households of about 2,000 to 6,000. They are built from groups of LSOAs and constrained by the local authority boundaries used for 2011 Census outputs. In the 2011 Census, there were 7,201 MSOAs in England and Wales. This increased to 7,264 from the 2021 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 in our Average household income, UK: Financial year ending 2018 bulletin.

The definitions of income and data sources used for these statistics are different. The Small Area Income estimates come from the Department for Work and Pensions' (DWP) Family Resources Survey, while those from other Office for National Statistics (ONS) publications come from the Household Finance Survey (FRS). 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. Such checks are described in Section 5 and include an assessment of residuals compared with model estimates, estimates of precision, stability and a Wald-based comparison of the direct survey and modelled MSOA estimates. 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 2020 with that used for financial year ending 2018 and guidance on the use of the estimates.

The methodology uses a range of admin and survey data sources including the Department of Work and Pensions' Family Resources Survey.

Further technical information, more background to the need for income estimates for small areas, and other methods considered are included in our previously published methodology, Income estimates for small areas technical report: financial year ending 2016.

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, survey sample 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. So, 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 our previously published methodology, Income estimates for small areas technical report: financial year ending 2016.

3 . Modelling for income, datasets

Survey data

The survey data were obtained from the Family Resources Survey: financial year 2019 to 2020, published on the GOV.UK website.

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)

Note that the definition of "equivalisation" is provided later in this section and considers the household size and composition. It acknowledges that, for example, two people do not need double the income of one person to have the same living standards.

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, and social 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, and 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 equivalisation.

Net annual household income after housing costs (equivalised)

This uses the same elements as net annual household income, but also deducts housing costs, such as 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. Equivalised income 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) equivalised income measures. For more details on these income definitions and the equivalisation scale, see our previously published methodology, Income estimates for small areas technical report: financial year ending 2016.

As was the case in the financial year ending 2018, we have published the financial year ending 2020 income 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 are modelled using weekly income data as per previous outputs. The final weekly estimates are expressed as annual income using a factor of 365.25 divided by 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 the financial year ending 2020, 28 addresses per PSU were selected. More information on the FRS methodology is contained within the FRS Background note and methodology report on the GOV.UK website.

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 the financial year ending 2020, the achieved sample size (for the UK) was 19,244 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 14,408 households from 1,170 postcode sectors in financial year ending 2020. The final survey data file for England and Wales contained cases in 2,551 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 33 (the maximum number of sampled households).

Consistent with the analyses for previous publications, for each different income type, a minority of records (302 of 14,408 - 2% for total annual household income) 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. 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.

For the net weekly (unequivalised and equivalised) income, records were removed where the net income was greater than the total income. 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, Households below average income data (HBAI), published on GOV.UK. This is a record-level dataset maintained by the Department for Work and Pensions.

Definitions from Family Resources Survey data

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

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 published on GOV.UK. 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 covariates 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. They are:

- Census 2021 - a wide range of variables relating to the MSOA where each FRS respondent is located; examples include the proportion of adults involved in managerial and professional work and the proportion of households who are defined as deprived in terms of health dimension.

- Department for Work and Pensions benefit claimant counts, August 2019 (provided as counts; see subheading DWP data)

- Valuation Office Agency (VOA) Council Tax Bandings, March 2019 (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) 2020 (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, 2019

- Her Majesty's Revenue and Customs, Pay as You Earn data, 2019

- Regional or country identification variable (for more information see subheading Regional or country identification variable)

Department for Work and Pensions data

The Department for Work and Pensions (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-2019 were used as denominators to derive these proportions.

Valuations Office Agency Council Tax bandings

The Valuations Office Agency (VOA) assigns each residential property in England to one of eight Council Tax bands, depending on its value on 1 April 1991. In Wales, each property is assigned to one of nine Council Tax bands depending on its value on 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.

Regional or country identification variable

England is split into nine ITL Regions Level 1. 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. The region and country variables included in modelling income were:

- North East
- North West
- Yorkshire and The Humber
- East Midlands
- West Midlands
- East of England
- South East
- South West
- Wales

Note that London was selected as the base case and therefore not specified separately in the modelling procedure.

The data used are as close to the reference period of the target income estimates as possible (that is, for financial year ending 2020). This is illustrated in Section 6: Comparing results for financial years ending 2018, 2020 and measuring change. Administrative data are collected primarily for government administrative processes and may change over time.

Data preparation

Before any modelling could proceed, substantial effort had to be channelled into gathering the necessary source data, principally survey response data and covariates 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.

While previous Small area income estimates releases (for example 2018) used the 2011 Census data, this release uses the new 2021 data. Consequently, almost all of the census covariates were updated. There existed 7,201 MSOA units in England and Wales in 2011. This increased to 7,264 in 2021.

For consistency with previous releases and the structure of the other covariates from 2019 to 2020, the 2011 MSOAs were reconstructed in terms of census data from the 2021 MSOAs. This meant that we could obtain the census profiles for each 2011 MSOA so that their contained census profiles would be up to date for 2021. This was carried out by weighting according to the numbers of postcodes contain in each MSOA. Such transitions only affected a minority of MSOAs whose borders had moved between the censuses.

The resulting matched dataset, containing the survey variable along with associated covariates and MSOA and Postcode Sector (the latter being the FRS Primary Sampling unit) 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.

As with the modelling for previous publications, where missing values existed for any of the covariates, the England and Wales mean of the variable in question was used to impute the missing value.

4 . Developing the models

Linear models were developed for England and Wales. These models take into account the fact that each individual household belongs to a specific area. They also 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 were fitted as multilevel models and can be used to produce estimates of the target variable at the small-area level. These 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" was 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 was 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 stepwise 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, statistically significant (at the 5% level) covariates were selected using a stepwise method 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 that covariates were sometimes included in the model even though they did not maintain significance at the 5% level once the interactions terms were included, since they were included in an interaction term, which is significant.

After modelling, adjustments are made to the modelled estimates to ensure they were consistent with the direct survey estimates at regional level for England and country level for Wales (this is known as "calibration"). The Family Resources Survey (FRS) 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 were aggregated to 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 was used to scale all of the modelled MSOA-level estimates and their confidence intervals. More details on this calibration and benchmarking methodology and aspects of the modelling methodology are given in our previously published methodology,Income estimates for small areas technical report: financial year ending 2016

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

Table 1: Key to covariates included in the model for total weekly household income, unequivalised

Covariate Name	Label	Source	T ratio
northest	Respondent is in North East	Country/regional indicators	0.25
northwst	Respondent is in North West	Country/regional indicators	0.82
york	Respondent is in Yorks / Humber	Country/regional indicators	-0.06
eastmid	Respondent is in East Midlands	Country/regional indicators	-0.64
westmid	Respondent is in West Midlands	Country/regional indicators	-0.40
east	Respondent is in East of England	Country/regional indicators	1.61
wales	Respondent is in Wales	Country/regional indicators	-1.44
southest	Respondent is in South East	Country/regional indicators	1.64
southwst	Respondent is in South West	Country/regional indicators	1.14
ewp14g0md	Standardised P14 - PAYE - All - Median	Admin	5.19
ewlnp14g5lq	Standardised Logit of P14 - PAYE - Females 60-64 - Lower Q'ile	Admin	1.75
lnisa2	Standardised Logit of Income Support - Age - 25-49; = of age2549 in population	Benefits_data	-1.84
phrpman	Proportion of HRPs aged 16 to 74 whose NS-SEC is 'managerial and professional'	Census	2.07
lnpborneur	Logit of Proportion of people born in Europe	Census	2.91
lnphhtype1	Logit of Proportion of households that contain one person only	Census	-3.68
lnengdef	Centrered Logit of propn of Band DEF England	Council Tax	3.06
ewp14g6lq	Standardised P14 - PAYE - Females 65+ - Lower Q'ile	Admin	1.80
lnuctot	Standardised Logit of Universal Credit - Total - Total = of over16 in population	Benefits_data	-1.54
lnpborneur_southest	Interaction of Logit of Proportion of people born in Europe with Respondent is in South East	Census & Country /regional indicators	2.80
lnuctot_york	Interaction of Standardised Logit of Universal Credit - Total - Total = of over16 in population with Respondent is in Yorks / Humber	Benefits_data & Country/regional indicators	-2.29
lnphhtype1_southest	Interaction of Logit of Proportion of households that contain one person only with Respondent is in South East	Census & Country /regional indicators	-2.24

Source: Office for National Statistics

With no covariates included in the model, the estimated standard residual area variance was 0.0463 (0.0041) compared with 0.0065 (0.0024) when the significant covariates are included in the model; a decrease of 86.01%. Therefore, these covariates together account for 86.01% of the total between area variance.

The most significant covariate in the model is the census covariate "ewp14g0md", which has a t-value of 5.19. As one would expect, this covariate has a positive coefficient: as median PAYE P14 income across all groups in an MSOA increases, so does the average household income for that MSOA.

"lnphhtype1" is the next most significant covariate in the model with a positive coefficient. This has a t-value of negative 3.68. Being negative, it shows that as the proportion of households that contain one person only increases, average household income decreases.

The relationship of a covariate with the average household income may be different if it is also involved in a model interaction. For example, the interaction variable "lnpborneur_southest" was found to be statistically significant. This suggests that the relationship between "lnpborneur" (the logit of the proportion of people born in Europe) and the average household income is different for MSOAs in the South East as compared with the rest of England and Wales. As the coefficient for both "lnpborneur" and "lnpborneur_southest" are both positive, this implies that a unit increase in the log of the proportion of people born in Europe has a greater effect in the South East than it has elsewhere in England and Wales.

Table 2: Key to covariates included in the model for net weekly household income (unequivalised)

Covariate Name	Label	Source	T ratio
northest	Respondent is in North East	Country/regional indicators	1.04
northwst	Respondent is in North West	Country/regional indicators	1.38
york	Respondent is in Yorks / Humber	Country/regional indicators	2.01
eastmid	Respondent is in East Midlands	Country/regional indicators	1.31
westmid	Respondent is in West Midlands	Country/regional indicators	1.54
east	Respondent is in East of England	Country/regional indicators	2.88
wales	Respondent is in Wales	Country/regional indicators	0.41
southest	Respondent is in South East	Country/regional indicators	2.73
southwst	Respondent is in South West	Country/regional indicators	2.47
lnphrpman	Logit of Proportion of HRPs aged 16 to 74 whose NS-SEC is 'managerial and professional'	Census	1.93
lnphhtype1	Logit of Proportion of households that contain one person only	Census	-5.36
pborneur	Proportion of people born in Europe	Census	3.54
ewlnp14g0mn	Standardised Logit of P14 - PAYE - All - Mean	Admin	3.64
lnengdef	Centrered Logit of propn of Band DEF England	Council Tax	2.37
ewlnp14g6lq	Standardised Logit of P14 - PAYE - Females 65+ - Lower Q'ile	Admin	2.56
ewp14g5md	Standardised P14 - PAYE - Females 60-64 - Median	Admin	2.26
lnwalghi	Centrered Logit of propn of Band GHI Wales	Council Tax	2.06
ewtcg3tp	Standardised P14 - PAYE - Males 65+ - 10th P'ile	Admin	2.00
lnengdef_pborneur	Interaction of Centrered Logit of propn of Band DEF England with Proportion of people born in Europe	Council Tax & Census	2.22
pborneur_northwst	Interaction of Proportion of people born in Europe with Respondent is in North West	Census & Country /regional indicators	1.94

Source: Office for National Statistics

With no covariates included in the model, the estimated residual area variance was 0.0318 (0.003) compared with 0.0048 (0.002), when the significant covariates were included in the model; a decrease of 84.84%. Therefore, these covariates together accounted for 84.84% of the total between area variance.

The most significant covariate in the model is the census covariate "lnphhtype1", (the logit of the proportion of households that contain one person only), which has a t-value of negative 5.36. As one would expect, this covariate has a negative coefficient; as the proportion of households that contain one person only increases, the average household income for that MSOA decreases.

The standardised logit of the mean P14 PAYE income across all persons in the MSOA, "ewlnp14g0mn", is the next most significant covariate in the model and has a positive coefficient and a T-value of 3.64. This also shows that, as the MSOA mean of the P14 PAYE income across all groups increases, so does the average household income.

Table 3: Key to covariates included in the model for net weekly household income, equivalised, before housing costs

Covariate Name	Label	Source	T ratio
northest	Respondent is in North East	Country/regional indicators	0.10
northwst	Respondent is in North West	Country/regional indicators	2.12
york	Respondent is in Yorks / Humber	Country/regional indicators	1.96
eastmid	Respondent is in East Midlands	Country/regional indicators	-1.99
westmid	Respondent is in West Midlands	Country/regional indicators	1.12
east	Respondent is in East of England	Country/regional indicators	3.25
wales	Respondent is in Wales	Country/regional indicators	-0.63
southest	Respondent is in South East	Country/regional indicators	3.21
southwst	Respondent is in South West	Country/regional indicators	2.80
ewp14g6lq	Standardised P14 - PAYE - Females 65+ - Lower Q'ile	Admin	3.28
ewlnp14g0mn	Standardised Logit of P14 - PAYE - All - Mean	Admin	7.10
lnisa4	Standardised Logit of Income Support - Age - 60+ = of over60 in population	Benefits_data	-2.48
lnpcts	Standardised Logit of Pension Credit - Total - Savings Credit only = of over16 in population	Benefits_data	-2.51
lnpecactiv	Logit of Proportion of people aged 16 to 74 who are economically active	Census	3.59
lnengdef	Centrered Logit of propn of Band DEF England	Council Tax	3.75
wabc_l	Centrered propn of Band ABC Wales	Council Tax	-2.42
lnlpwktcp	Centrered Logit of propn of Total lone parents families as propn of adult popn	Tax_credits_data	2.85
lnengdef_lnpcts	Interaction of Centrered Logit of propn of Band DEF England with Standardised Logit of Pension Credit - Total - Savings Credit only = of over16 in population	Council Tax & Benefits_data	-2.48
lnpecactiv_eastmid	Interaction of Logit of Proportion of people aged 16 to 74 who are economically active with Respondent is in East Midlands	Census & Country/regional indicators	2.17
ewlnp14g0mn_york	Interaction of Standardised Logit of P14 - PAYE - All - Mean with Respondent is in Yorks / Humber	Admin & Country /regional indicators	2.15

Source: Office for National Statistics

With no covariates included in the model, the estimated residual area variance was 0.0253 (0.002) compared with 0.0028 (0.001) when the significant covariates were included in the model, a decrease of 88.91%. Therefore, these covariates together accounted for 88.91% of the total between area variance.

The most significant covariate in the model is the census covariate, "ewlnp14g0mn" which has a t-value of 7.10. As the mean for an MSOA of the P14 PAYE income across all groups increases, so does the income.

The next most significant covariate is "lnengdef" in the model with a positive coefficient and a t-value of 3.75. This shows that, as the centred Logit of the proportion of Band DEF properties in English MSOAs increase, the average household income increases.

The most significant variable with a negative coefficient was "lnpcts" (the standardised logit of the proportion of people aged 16 years who receive Pension -- Savings Credit), which had a t-value of negative 2.51. This shows that as the proportion of claimants for the above increases, the average household income for that MSOA decreases.

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

Covariate Name	Label	Source	T ratio
northest	Respondent is in North East	Country /regional indicators	-0.45
northwst	Respondent is in North West	Country /regional indicators	1.83
york	Respondent is in Yorks / Humber	Country /regional indicators	0.92
eastmid	Respondent is in East Midlands	Country /regional indicators	0.74
westmid	Respondent is in West Midlands	Country /regional indicators	0.66
east	Respondent is in East of England	Country /regional indicators	2.01
wales	Respondent is in Wales	Country /regional indicators	-0.41
southest	Respondent is in South East	Country /regional indicators	1.97
southwst	Respondent is in South West	Country /regional indicators	1.54
ewlnpayeg0lq	Standardised Logit of Self Assessment - All - Lower Q'ile	Admin	-2.65
ewp14g4mn	Standardised P14 - PAYE - Females 16-59 - Mean	Admin	3.44
ewp14g6lq	Standardised P14 - PAYE - Females 65+ - Lower Q'ile	Admin	3.29
lnisa4	Standardised Logit of Income Support - Age - 60+ = of over60 in population	Benefits_data	-2.50
lnjsam	Standardised Logit of Job Seekers Allowance - Gender - Male = of malework in population	Benefits_data	-2.49
ewaaa2	Standardised Attendance Allowance - Age - 70+ = of over70 in population	Benefits_data	-2.27
phhdepr_hous	Proportion of households classed as deprived (housing)	Census	-5.49
lnphrpman	Logit of Proportion of HRPs aged 16 to 74 whose NS-SEC is 'managerial and professional'	Census	3.09
edef_c	Standardised raw counts of Band DEF England	Council Tax	2.50
edef_c_ewlnpayeg0lq	Interaction of Standardised raw counts of Band DEF England with Standardised Logit of Self Assessment - All - Lower Q'ile	Council Tax & Admin	-2.08

Source: Office for National Statistics

With no covariates included in the model, the estimated residual area variance was 0.0304 (0.003) compared with 0.0029 (0.002) when the significant covariates were included in the model, a decrease of 90.38%. Therefore, these covariates together accounted for 90.38% of the total between-area variance.

The most significant covariate in the model is the census covariate '"phhdepr_hous", which has a t-value of negative 5.49. As the proportion of households classed as deprived (housing) within the MSOA increases, the average household income for that MSOA decreases.

The most significant positive covariate in the model is "ewp14g4mn", with a positive coefficient and a t-value of 3.44. This shows that, as the standardised MSOA mean (P14) PAYE earnings recorded for females aged 16 to 59 years increase, so does the average household income.

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:

1. Across the four models, covariates relating to PAYE amounts are one of the strongest determinants of household income at small area level. PAYE, as administrative sources of data, are particularly useful as they are available for a high proportion of individuals across the country. Household size is important as a negative correlator for unequivalised incomes with single-person households earning lower incomes than larger ones.

2. Other significant positive drivers of income at MSOA level include the proportion of Europe-born residents and the proportion of higher value (Council Tax bands D, E and F) properties. Negative drivers include the proportion of households classed as deprived (housing) and the proportion of people taking Pension Credits.

3. The majority of regional or country indicators in each model are not significant but are included. The reason for this is that balancing ("calibration") is carried out on the raw income estimates to balance regional and Wales-level average income estimates to those directly derived from the Family Resources Survey.

4. The final types of covariates included in the models are interaction effects. The majority of the interaction terms involve regional or country indicators. This shows that some covariates have different effects in different regions.

Some of the 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 as described in this section, in order 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 and published in the article, Evaluation of small area estimation methods - An application to unemployment estimates from the UK LFS, on the Research Gate website, 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 the confidence intervals produced from the models.

The results of the diagnostics for all four income types are summarised in Table 5, with further information about each given after the table. More detail about the diagnostic tests and why they are performed can be found in our previously published methodology, Income estimates for small areas technical report: financial year ending 2016.

Table 5: 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 Income Equivalised Before Housing costs	Net Weekly Household Income Equivalised After Housing costs
Residual vs Model Estimates	**Constant (SE)**	-0.831 (0.205)	-0.880 (0.214)	-0.635 (0.198)	-0.508 (0.200)
Household Level Residuals	**Slope (SE)**	0.123 (0.030)	0.136 (0.033)	0.099 (0.031)	0.080 (0.032)
Residual vs Model Estimates	**Constant (SE)**	-0.127 (0.026)	-0.124 (0.025)	-0.084 (0.019)	-0.045 (0.015)
Area Level Residuals	**Slope (SE)**	0.019 (0.004)	0.019 (0.004)	0.013 (0.003)	0.007 (0.002)
Model vs Sample Estimates	**Constant (SE)**	-75.507 (40.542)	-111.94 (38.76)	-81.299 (31.605)	-17.838 (27.996)
	Slope (SE)	1.068 (0.045)	1.150 (0.057)	1.129 (0.050)	1.037 (0.049)
Model vs Sample Estimates	**Constant (SE)**	-25.978 (132.75)	122.47 (164.70)	-19.499 (128.105)	-65.937 (99.699)
	Slope (SE)	0.960 (0.278)	0.465 (0.471)	0.933 (0.396)	1.206 (0.341)
	Quadratic term (SE)	0.00006 (0.0001)	0.0005 (0.0003)	0.0001 (0.0003)	-0.0001 (0.0003)
Coverage	**%**	100	99.96	100	100
Wald	**P-value**	1	1	1	1
Stability Analysis	**RRMSE**	0.051	0.044	0.035	0.046

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 our previously published methodology, Income estimates for small areas technical report: financial year ending 2016 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 may have 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 zero.

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 equals 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 equals x" line.

An important assumption when using this diagnostic is that the direct estimates are unbiased. The technique for calculating direct survey estimates at an MSOA level is described in our previously published methodology, Income estimates for small areas technical report: financial year ending 2016, 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 zero and the slope term is not significantly different from one. Therefore, the fit is very close to the "y equals x" line. This shows that at least in sampled areas, the modelled estimates do show either none or only occasional very small signs of bias; ones that are in line with previously published results.

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 our previously published methodology, Income estimates for small areas technical report: financial year ending 2016. 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 of 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 into two datasets similar in size and MSOA representation. 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. The RRMSE stability measures for the four models are all low, ranging from 0.035 to 0.051, thus indicating a high degree of stability; generally higher still than the models for FYE 2018 (which ranged from 0.049 to 0.072) indicate that the different sets of data produce similar sets of estimates.

6 . Comparing results for financial years ending 2018, 2020 and measuring change

Middle-layer Super Output Area (MSOA)-level model-based estimates of average annual household income have been produced for financial year 2020 in England and Wales, fulfilling users' requirements for income information at MSOA level.

Models

In financial year ending 2018, each model related the Family Resources Survey (FRS) survey estimate of weekly household income to 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 (i. e: Higher and intermediate managerial, administrative, professional occupations
- Proportion of people living in communal establishments

In financial year ending 2020, the only variables that were part of all four models were region and country and these were included by design. However, there were two other variables that were included in three of the four models:

- centred Logit of the proportion of Council Tax bands D, E and F properties in England
- standardised MSOA lower quartile (P14) PAYE earnings recorded for females aged 65 years and over

Diagnostics

Some plots of household- and area-level residuals for all models, for both financial years ending 2018 and 2020, 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 pattern.

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 2018 and 2020 models produce fairly consistent results indicating that in general the models for England and Wales are well-specified and the assumptions are satisfied. The percent of variability explained in the Gross and Net unequivalised income models for 2020 exceeded those for 2018. For net equivalised before housing costs, this remained as strong in 2020 as in 2018. For after housing costs, this did not reach the very high level seen in 2018, but was still strong. This demonstrates 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 year ending 2018:

- Census data, 2011
- HMRC PAYE data, March 2018
- DWP benefit data, 2017
- Region/country indicators
- ONS House Price Statistics for Small Areas year ending March 2017
- Council Tax data, March 2017
- DECC Energy Consumption data 2017

The following covariate data were used in the model-based estimates of income for the financial year ending 2020:

- Census data, 2021
- HMRC PAYE data, March 2019
- DWP benefit data, August 2019
- Region/country indicators
- ONS House Price Statistics for Small Areas year ending March 2020
- Council Tax data, March 2019
- DECC Energy Consumption data 2019

These lists of data sources show that different covariate datasets were available and used at the time of modelling the financial years ending 2018 and 2020 model-based estimates of average income.

Different covariates have been selected in the models for financial years ending 2018 and 2020. 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 2018 and 2020 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.

As mentioned in Section 3: Modelling for income, datasets, the 2021 Census data were extracted for each of the published 7,264 MSOAs. For consistency with previous releases, estimates of compositions for each of the 7,201 MSOAs in the previous set were made and used in the modelling.

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, 2018 and 2020 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, 2018 and 2020 estimates use the 2011 Census geography, which contains more MSOAs and some altered MSOA boundaries compared with previous estimates, which were based on the 2001 Census geography. Therefore, for some MSOAs, direct comparison 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 2020 (or 2018, 2016, 2014 or 2012), will not match with any geography code for earlier estimates. Note that the 2020 estimates were different from earlier ones in that they were modelled from 2021 rather than 2011 Census data. However, the 2020 income estimates for MSOAs are comparable to their 2018, 2016 and 2014 counterparts in this respect as exactly the same MSOA geography and borders were applied.

If the confidence intervals for the estimates at different time periods do not overlap, 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.

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 2020 survey data. As such, 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 statistically 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, an 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 2011 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 2020 survey data and so are only valid for this period.

8 . Cite this methodology

Office for National Statistics (ONS), released 11 October 2023, ONS website, methodology, Income estimates for small areas in England and Wales, technical report: financial year ending 2020