United Kingdom

Office for National Statistics

https://www.ons.gov.uk/

SAE motivation

Data on income at the smallest possible geographical level helps with identification of communities that have large levels of deprivation or inequality, and as such can be used to support work addressing issues of inequality. To gain this information, we need answers to some questions such as the following:

What is the average household disposable income in different local areas in England and Wales?

Where are the highest and lowest income areas located?

What is the average income adjusted for housing costs in different areas?

Including questions on income information in the England and Wales Census has been considered regularly. However, due to concerns about the public acceptability of asking about their income, it has never been incorporated in the Census. This concern was confirmed during a Census test conducted in 2007 that contained a question on income, leading to lower response rates. **Instead, small area estimation methods have been implemented to produce official income estimates at the small area level for England and Wales.**

Indicators in the scope of the study and Input data

The Family Resources Survey (FRS) is the largest survey available that collects information on household income. It covers around 19,000 to 20,000 private UK households and through questionnaires and interviews asks respondents questions on income and its components, including benefits, tax pensions and tax credits. Based on FRS response, four different measures of household income are derived for use in modelling:

- **Total household weekly income (unequivalised)**

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

- **Net household weekly income (unequivalised)**

This is calculated as total household weekly income but is net of income tax payments, national insurance contributions, domestic rates/council tax, contributions to occupational pension schemes, all maintenance and child support payments, which are deducted from the person making the payments, and parental contribution to students living away from home.

- **Net household weekly income before housing cost (equivalised)**

This is calculated in the same way as net household weekly income but with the application of OECD’s equivalisation scale to adjust the household income values to represent the income level of every individual in the household.

- **Net household weekly income after housing costs (equivalised)**

This is calculated in the same way as equivalised net household weekly income before housing cost but is subject to the deduction of rent, water rates, community water charges and council water charges, mortgage interest payments, structural insurance premiums (for owner occupiers), ground rent and service charges.

All variables above are obtained from FRS, however net equivalised household weekly income before and after housing costs is defined and calculated based on Households Below Average Income (HBAI) methodology. For further detail on FRS or HBAI please see a note on Family Resources Survey: methodology and background notes.

The following additional data sources were available during the model selection process, but different years’ estimates included a different subset in the final estimation model due to availability and model selection:

- Census
- Department for Work and Pensions benefit claimant counts
- Valuation Office Agency (VOA) Council Tax Bandings
- Office for National Statistics, House Price Statistics for Small Areas
- Department of Energy and Climate Change, Energy Consumption data
- Her Majesty’s Revenue and Customs, Pay As You Earn (PAYE) data
- Regional or country identification variable

Many of these sources are administrative data sources that have not been collected for statistical purposes and thus have differing coverage and definitions, especially compared with the FRS. Nevertheless, they are thought to include relevant information associated with income and could be potentially useful predictors for estimates at an area level. The list of alternative data sources available for the potential inclusion into the model is growing with more data sharing across departments.

Small area estimation modelling techniques enable us to build models that take their strengths from survey, census and administrative data sources. The final dataset for modelling is a combination of different data sources with survey household records linked to auxiliary data from census and administrative data, available at area level, using postcode variable as a match-key.

Data challenges

The survey is designed to provide direct estimates of good quality at national and regional levels, but no levels lower than that. The sample is too small to provide reliable direct estimates for small areas. As such, model-based methods are required for small areas. These are based on model parameters and values for the covariate (auxiliary) data. Furthermore, only 14% of areas (postcode sectors) were sampled in the latest income estimates (tax year ending 2018). The Great Britain FRS uses a stratified clustered probability method and samples 1,417 postcode sectors from 9,200. Using sampling techniques may lead to bias or sampling error. The small area modelling approaches include random effects to account for the clustering.

SAE model building

SAE methods/Specification

The first stage is to determine independent variables from census and administrative sources that are relevant to each of the four variables of interest (dependent variables). Forward and backward selection is used to help identify significant covariates to be included, with region/country indicator terms forced into the model. Selected covariates and two-way interactions are retained for estimation. This step is rerun for each new publication to ensure the relevance of data sources over time. However, changes in the covariate selection over time limit conclusions on associated estimates of change as any observed changes could be attributed to changes in covariates over time to some degree.

Once covariates have been determined for each dependent variable of interest, multilevel linear models were used to produce Middle Layer Super Output Area (MSOA) level estimates of mean household income for four dependent variables: average weekly gross and net household income, average weekly net household income (equivalised), and average weekly equivalised net household income after housing costs. Weekly household income was used as the dependent variable and the area level covariates as explanatory variables. The models relate the household-level survey variable to the covariates that relate to the small area where that household is located. Although the model outputs provide MSOA level mean estimates, model inputs are at household level**.** Non-linear estimates, like median and percentile, cannot currently be derived in this way. The MSOAs discussed in this report are based on Census data from 2011, and they have a mean population of 7200 and minimum of 5000.

Statistical software

The multilevel models were fitted using SAS, which was the office standard software at the time of the initial model development. Development of a pipeline to convert the code from SAS to R is currently being explored.

Model validation and benchmarking

Diagnostic checks performed on the models include the following:

- Residuals plotted against modelled estimates – at household and at area (postcode sector) level. Ideally the residuals should have a constant variance across the modelled estimates. If there is a pattern in the residuals, the model may be mis-specified (an important covariate has been left out of the model). This diagnostic also tells us whether there is non-constant variance of the residuals.
- Comparison of modelled estimates to direct survey estimates – this diagnostic plot tells us whether the modelled estimates are unbiased. Ideally, the regression line between the modelled and survey estimates should be close to x=y. If this is not the case (i.e. the regression line is far from the x=y, or is curved), then the model may be mis-specified.
- Coverage diagnostic – this is used to check whether there is sufficient overlap between the confidence intervals for the survey estimates compared with the confidence intervals of the modelled estimates. Ideally, at least 95% of MSOA's (where there is a survey estimate) should have overlapping confidence intervals from the survey and the model estimate.
- Distinguishability – this diagnostic determines the percentage of MSOA's with lowest modelled estimates that have confidence intervals overlapping with confidence intervals from MSOA's with the highest modelled estimates. As a practical guide it is considered that, at least 20% of confidence intervals at the lower range should not overlap with 20% of confidence intervals at the upper range.
- Wald statistic – this is a goodness-of-fit statistic to test whether the modelled estimates are significantly different from the survey estimates. There should ideally be no significant difference assuming there is a close relationship between the modelled and survey estimates.
- Stability analysis – this diagnostic assesses the predictive power of the model. For this, the survey data is split into two sub-datasets (A and B). For both sub-datasets, the income model is re-run on each to derive two sets of regression coefficients. Both sets of regression coefficients are used to produce modelled estimates for all MSOAs. The modelled estimates from sub-datasets A and B are compared by calculating the relative root mean square error (RRMSE). This procedure is repeated for 10 runs. Model stability is assumed if the median RRMSE value across the 10 runs is less than 0.5.
- Coefficients of variation – This diagnostic tells us whether the modelled estimates are of an acceptable quality of precision. For each MSOA, the Coefficient of Variation (CV) is calculated by dividing the modelled estimate by its standard error. CVs are normally acceptable if they are less than 20%, though at MSOA level CV values may be accepted up to 30%.

These diagnostics are used to refine the models where selection of covariates may be varied, for instance to see if changing covariates improves model diagnostics.

Model validation is performed by making comparisons of direct estimates of income and model-based estimates of income at MSOA level.

Benchmarking of the data is done to ensure that the MSOA estimates are consistent with the regional survey estimates. This is used to avoid inconsistencies between the model estimates and direct survey estimates. Direct survey estimates of income were calculated at the region (in England) and country (Wales) levels from the FRS data. Similarly, the MSOA model estimates were aggregated to derive regional totals. A scaling factor is calculated by dividing the survey regional estimates by the sum of the model-based estimates for that region. The model-based estimates are multiplied by this scaling factor. The Technical report provides further information.

External review process

As the small area income estimates are designated as a National Statistic to retain such a status, a semi-regular review of the collection, production and dissemination of such data occurs. There is currently a compliance review being undertaken with the Office for Statistical Regulation. The production team plans to undertake further engagement with current, previous and potential users of this data to see what elements are of most value, what improvements and developments they would see as beneficial, and to further communicate differences in comparison and coherence that may occur when comparing with other sub-national income estimates such as the admin-based estimates above, and National Accounts-benchmarked disposable household income estimates.

Modelling limitations/ challenges

There are some limitations of the current blended survey/administrative data modelling approach, some of which cannot be improved while others may require wider developments. For example:

- The production of small area estimates is still ultimately constrained by survey (FRS) sample size, or more specifically, the coverage for the geographical units of interest. Improving this challenge would be through a costly process of increasing the sample size.
- The modelling process tends to shrink estimates towards the average level, especially the higher and lower incomes at either end of the distribution.
- The four income types are separately modelled. As such, there may be inconsistencies when comparing different types of income (e.g. gross vs net income) in some areas. A more sophisticated model explicitly accounting for the relationship between some of the different income types may be possible but with further research required.
- The level of estimation is not the same as the level of clustering in the survey data, random mixed effects are used to compensate for this but the variance of household income at the postcode level is taken to be representative of the variance of household income at MSOA level in calculating confidence intervals. The technical report provides further information. This cannot be solved under the current sources used.

Non-linear estimates, like the median and percentile, cannot be derived using the existing model specification. Using the mean provides only a summary measure of household income and does not capture more detailed information about the distribution. Recognising user needs for non-linear estimates, ONS has previously explored alternative methods to derive them. Please note this research is experimental in nature and therefore has not gone through a rigorous approval process like ONS National Statistics-designated outputs.

Estimating distributions of household income for middle layer super output areas in 2011 using small area estimation methods - Office for National Statistics

Link to the latest small area income estimates/report: Small area model based income estimates 2018

Future development

ONS is considering whether elements of these methods could add value to the experimental admin-based income statistics (ABIS). ABIS are part of a progr amme for the future of population and social statistics that aims to provide more frequent, relevant and timely statistics that will better meet users' needs to understand the population and how it changes. They are produced using data from HM Revenue and Customs' (HMRC) Pay As You Earn (PAYE), Self-Assessment and the Department for Work and Pensions' (DWP) benefit systems. The outputs consist of percentiles of gross and net income for individuals and households at Lower Layer Super Output Area level. The outputs for household income refer to income at an address which is different to the traditional household definition used for the model-based statistics. The differences between these household definitions are discussed in the Occupied address research output. As the admin-based income statistics are still under development, they do not currently capture all required components of income, meaning the income measures and coverage are incomplete. As such they have limited use for decision-making at this time. For further information see the latest release Admin-based income, England and Wales.

SAE work within the organisation

The output production teams in ONS have topic area expertise and run the estimation for new publications. The Small Area Estimation (SAE) expert group in the Methodology and Quality Division (MQD) of ONS has a long-standing experience in Small Area Estimation and the model development for income. The SAE expert group provides a formal quality assurance prior to the publication of new estimates covering model fitting, model performance and the quality of the estimates obtained. This quality assurance together with user feedback on the estimates builds evidence of the fitness for purpose of the outputs and is embedded in their classification as National Statistics.

Information provided by The Small Area Estimation Team at the Office for National Statistics.