Office for National Statistics

Model-Based Estimates of households in poverty for Middle Layer Super Output Areas, 2011/12 Technical Report

Office for National Statistics

Official Statistics

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

About us

The Office for National Statistics

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

The Government Statistical Service

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

Contacts

This publication

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

Other customer enquiries

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

Media enquiries

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

Copyright and reproduction

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

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

Table of Contents

Executive summary

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

This report provides technical information about the methods and processes used to produce the middle layer super output area (MSOA) estimates of the proportion of households in poverty in England and Wales:

• MSOA level estimates of the proportion of households below 60 per cent of the UK median income after housing costs (AHC);
• MSOA level estimates of the proportion of households below 60 per cent of the UK median income before housing costs (BHC);

This follows the previous publication of the AHC estimates, for 2007/08. Household mean income estimates for 2013/14 are also being published following the previous publication for 2011/12, and these estimates are for the following four income types:

• total household weekly income (unequivalised);
• net household weekly income (unequivalised);
• net household weekly income before housing costs (equivalised); and
• net household weekly income after housing costs (equivalised).

These estimates are produced following the continued user need for estimates tackling more specific poverty measures, particularly following the HBAI indicator of those below 60 percent median (both for before and after housing costs).

In carrying out the modelling, the set of possible covariates considered were the same set that had been considered for mean income modelling for 2011/12. These were MSOA level indicators from the 2011 census, 2011 DWP benefit data, 2011 Council Tax data, 2011 HMRC Child Tax Credit and Working Tax Credit data, 2009 CLG change of ownership by dwelling price data and 2011 Council Tax Data. Region indicators were also included. Finally, interactions between covariates were considered. This set of covariates contains the primary variables expected to influence household poverty levels. Covariate selection was carried out in a controlled way by considering groups of variables. For example, variables from the 2011 Census or DWP data initially each go through a separate selection. Further refinements are then made based on the log likelihood and significance of covariates. The variables considered were selected on the basis that it was expected that these variables had some relationship with income.

A number of diagnostic checks are used to assess the appropriateness of the model and quality of the estimates. The checks show that modelling assumptions are satisfied.

A measure of the explanatory power or goodness of fit of the modelling is how much unexplained area level variability remains in the accepted model compared with what exists in a model with no explanatory covariates (the null model consisting of just the intercept). For 2011/12 (AHC) the percentage of between area variability explained by the chosen model is 82.3 per cent. For 2011/12 (BHC) the percentage of between area variability explained by the chosen model is 94.2 per cent. Although the AHC model has performed less well compared with the both the BHC and previous AHC models, the reduction in between area variability explained by the chosen model is not sufficiently large to cause concern about the overall quality of the model.

Another performance criterion usually adopted in assessing the publication quality of model-based estimates is distinguishability between areas. Because of the associated confidence intervals, areas cannot be judged as different simply on the basis of point estimates. With a very large number of areas, such as is the case for MSOAs, a simple (though conservative) test between two areas is to see if the confidence intervals overlap. If they do not then they can be judged significantly different. For the 2011/12 AHC model about 21 per cent of MSOAs at the lowest ranks can be statistically distinguished from the same number of the highest ranks (that is, their confidence intervals do not overlap). Similarly, for the 2011/12 BHC model about 27 per cent of MSOAs at the lowest ranks can be statistically distinguished from the same number of the highest ranks.

The analysis shows that both the AHC and BHC models proposed are well specified, performing well in terms of explaining between area variability, estimate precision and distinguishability between areas. The relationship between the poverty estimates and the published mean income estimates is also consistent. Therefore, these estimates from the models presented for 2011/12 are to be published as Experimental Statistics.

1. Introduction

Income information is needed at small area level in order to help identify deprived and disadvantaged communities and to support work on social exclusion. This requirement was previously reflected by Census User Groups who made a strong case for a question on income to be incorporated in the 2001 Census. Although this need was recognised by the government, concerns were also expressed about the sensitivity of an income question and, as a result, a question on income was not included in the 2001 Census and the same decision was taken for the 2011 Census. Instead, alternative methods for obtaining data on income at small area level were identified and implemented leading to the use of small area estimation methodologies to produce local area income estimates.

ONS has published these household mean income model based estimates for MSOA based on data from the FRS and Households Below Average Income (HBAI) statistics for 2004/05, 2007/08 and 2011/12 following previous publication at ward level in 1998/99 and 2001/02. Four measures of mean income have been published each time – household total gross income, household net income, equivalised household income before housing costs (BHC) and equivalised household income after housing costs (AHC).
Users (who have welcomed the publication of model based estimates of mean income) have also expressed the need for estimates of more specific poverty measures, such as the HBAI indicator of those below 60 per cent of the UK median household weekly income and preferably on a basis of persons or children. There are two such measures of poverty currently used by the DWP – AHC and BHC. DWP strongly support the need from users for producing small area poverty estimates.
ONS Methodology investigated the possibility of generating such estimates based on proportions of households (rather than persons) as a first step to the ideal basis, as the modelling more naturally integrates with the methodology developed for the mean estimation of the continuous quantity of household income itself.

The work undertaken resulted in the production of 2007/08 MSOA-level estimates of the proportion of households in poverty for England and Wales, calculated based on equivalised household income (AHC) and produced using the Small Area Estimation Programme (SAEP) methodology. This is the same methodology that was used to produce mean income estimates (http://www.ons.gov.uk/ons/rel/ness/small-area-model-based-income-estimates/2011-12/technical-report.pdf). The underlying small area estimation model uses unit level survey responses but area level covariates due to the well known restrictions to link unit level survey, census and administrative data.

Further information on the development of the 2007/08 poverty estimates can be found in the previously published 2007/08 Technical Report
(http://www.neighbourhood.statistics.gov.uk/HTMLDocs/images/TechnicalReport%20v1_0_tcm97-99412.pdf).

Super Output Areas 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 population size and whose boundaries do not change. This means fairer comparisons can be made between areas and over time. These areas are built from groups of Output Areas (OAs) used for the 2011 Census. OAs can be aggregated to form Lower Layer Super Output Areas (LSOA) and then the larger Middle Layer Super Output Areas (MSOA). MSOAs have a mean population of 7,800 and a minimum population of 5,000. In keeping with National Statistics geography and statistical policy for publishing statistics for Super Output Areas, ONS has produced the 2011/12 estimates of poverty at MSOA level. MSOAs are also comparable in size to the primary sampling unit (PSU) that the survey data are collected on. Therefore the estimates have better precision at MSOA level than at other levels, especially higher levels of geography such as LAD.

This report therefore presents results for 2011/12 MSOA level estimates of the proportion of households below 60 per cent of both median income AHC and BHC. The report is structured as follows: Section 2 describes the SAEP methodology in general and Section 3 describes its application to the problem of estimating proportions of households in poverty .Section 4 describes how the models were developed. The fitted model is described in Section 5 with the model based estimates being presented in Section 6. An assessment of the quality of the estimates is given in Section 7.

2. Small area modelling

The principal reasoning behind the need for small area estimation is that surveys are designed to provide reliable estimates at national and sometimes regional levels – they are not typically designed to provide estimates at lower geographical levels (for example local authorities and MSOAs). There is also the problem of sample design – with the exception of the Labour Force Survey (LFS) most of the principal national household surveys in UK have clustered designs. This means that the sample is not distributed totally randomly across the nation but that certain areas are first selected as PSUs (Primary Sampling Units) and then households are selected for interview from these. The areas selected as PSUs are usually postcode sectors. The selection of these is stratified in such a way that their distribution is nationally representative.

The problem for estimation at the small area level is that, irrespective of the total sample size, with clustering like this the inevitable result for areas such as MSOAs is that the majority will contain no sample respondents at all. Hence no direct survey estimates would be possible. Also, where there are estimates for particular MSOAs the sample sizes would be so small that the variability around the estimates would be too high for reliable estimates. MSOAs and PSUs are often of similar size in terms of households.

Following some preliminary studies into small area estimation, ONS Methodology established the SAEP in April 1998. The SAEP methodology involves combining survey data (in this case – income related variables/indicators) with other data that are available at the small area level and building a modelled relationship. The small area level would tend to be an area for which direct survey estimates cannot be produced due to their unreliability. The area level relationship between the survey variable and auxiliary variables (covariates) is estimated by regressing individual survey responses on area-values of the covariates.

In other words, the basic aim of the SAEP methodology is the construction of a statistical model relating the observed value of the survey variable of interest (measured at individual, household, or address-level) to the covariates that relate to the small area in which the address is located. These covariates are generally average values or proportions relating to all individuals or households in the area. They are generally administrative or census based, as they must have full coverage in all areas being modelled. Once the covariates have been selected and the model has been fitted, the model parameters can be applied to the appropriate covariate values for each and every area, hence obtaining estimates of the target (or survey) variable for all small areas.

While the model is constructed only on responses from sampled areas, the relationships identified by the model are assumed to apply nationally. 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 every area. This is the basis of the synthetic estimation ONS has already used to produce the estimates of average (mean) income for MSOA for 2004/05, 2007/08 and 2011/12.

2.1 Other approaches to poverty estimation

Other methods for estimating poverty include the World Bank method proposed by Elbers et al (2003), the Empirical Best Prediction (EBP) approach proposed by Molina and Rao (2010) and the M-quantile approach (Tzavidis et al, 2006, 2008, 2010). The World Bank approach involves fitting a model to the log of household per capita expenditure using survey data and using the fitted model to obtain estimates of expenditure for all areas using census covariates. The procedure to obtain estimates is repeated so that there are many replicas. Poverty indicators are calculated using each replica and then averaged.

The advantages of this method are that a unit level link between the survey covariates and census covariates is not required. Estimates of precision are available and there is ready to use software, POVMAP, available. However, there are also several disadvantages. The model fitting is ad-hoc and is not the usual multi-level modelling considered in small area estimation. Covariate selection is performed using preliminary regression models, not the assumed random effects model. There is no link between the covariates used to fit the model and those used to obtain the small area estimates.

The Empirical Best Prediction (EBP) approach fits a unit level random effects model using unit level covariates from census data to a suitable welfare measure, for example, income (obtained from a sample survey). The fitted model is used to obtain estimates for all areas since the covariates are from the census. An appropriate poverty measure is then obtained. This approach has several advantages over the World Bank method. The model fitting and selection is not ad-hoc and are carried out using standard procedures. Unit level covariates from the census are used to fit the model unlike in the World Bank method. Disadvantages of the EBP approach are that a unit level link between the survey and census is required and model assumptions include normality and non-informative sampling. The mean squared error (MSE) estimation is also highly computer intensive.

The M-quantile approach offers robust estimation of poverty indicators. This method models the quantiles of the distribution of a suitable welfare estimate and uses influence functions for estimation. This approach offers the advantage that estimation is robust to outliers in the data. Also normality is not a requirement. However this method is not yet fully developed.

This renewed interest in poverty estimation extends beyond the academic community with National Statistical Offices around the world showing interest in poverty mapping methodologies.
The methodology we employ is consistent with the methodology used for the previously published estimates of household poverty (2007/08) as well as estimates of mean income. Further development of the model and investigations into alternative approaches will be considered for future work.

3. Modelling the poverty indicator

3.1 Introduction

This section describes how the general SAEP methodology has been applied to the specific problem of estimating poverty at MSOA level. The datasets (both survey and covariate) used in the modelling process are described as well.

When summarising skewed data, such as income, alternative measures of the distribution are generally preferred over the mean or average as these statistics are sensitive to asymmetric distributions. The distribution of a continuous variable, such as income for a household, can be summarised using the median or percentiles. Traditionally the approach to measuring low income or poverty has been to look at how many people, households or families have an income that falls below some threshold. The threshold is commonly set at a particular fraction of mean or median income, calculated across the whole population, with 60 per cent of the median being a widely used threshold.

3.2 The data sets

3.2.1. The survey data

The survey data used in this modelling exercise come from the HBAI datasets that are prepared by DWP using data from the 2011/12 FRS (FRS report: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206887/frs_2011_12_report.pdf).
The FRS was chosen as the source for survey data for this study since it is the survey with the largest sample that includes suitable questions on income. The target parameters to be estimated are:

- the proportion of households below 60 per cent of the national median income based on net weekly household equivalised income AHC;
- the proportion of households below 60 per cent of the national median income based on net weekly household equivalised income BHC.

However, as the SAEP methodology uses household level responses, the survey variable to be modelled is a binary variable that indicates whether, when income is defined as net weekly equivalised income, the household's income lies below 60 per cent of the UK median income as reported in the HBAI report. The threshold values for 2011/12 are as published by DWP are (https://www.gov.uk/government/statistics/households-below-average-income-hbai-199495-to-201112):

1) £220 per week that corresponds to 60 per cent of UK median after housing costs equivalised net income (£367 per week)

2) £256 per week that corresponds to 60 per cent of UK median before housing costs equivalised net income (£427 per week).

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 in order to make sensible income comparisons between households.

These estimates use the OECD equivalisation scale. This was in response to the Government's 2004 Spending Review, which stated that future child poverty measurements will report incomes before housing costs and equivalised using the OECD scale. More information on the equivalisation scale is available in the HBAI report.
(https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206778/full_hbai13.pdf)

The FRS uses a stratified clustered probability sample drawn from the Royal Mail's small users Postcode Address File (PAF). The survey selects 1,848 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 2011/12, 24 addresses were selected per PSU. More information on the FRS methodology is contained within the FRS technical report
(https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206887/frs_2011_12_report.pdf)

The FRS aims to interview all adults in a selected household. A household is defined as fully co-operating when it meets this requirement. In addition, to count as fully co-operating, there must be less than 13 'don't know' or 'refusal' answers to monetary amount questions in the benefit unit schedule (i.e. excluding the assets section of the questionnaire). In 2011/12 the achieved sample size (for the UK) was 20,763 households.

The requirement for this project is to produce MSOA level estimates of the proportion of households in poverty for England and Wales. The survey data file used contained 15,493 households from 1,170 postcode sectors. The final survey data file for England and Wales contained cases in 2,587 different MSOAs out of a total of 7,201. The number of cases per MSOA in the achieved FRS sample varies widely particularly due to the fact that MSOAs cut across the postcode sectors, the primary sampling unit. For example, some MSOAs recorded only 1 response whereas, others had 32 (the maximum number of sampled households).

3.2.2. The covariate data sets

The small area estimation methodology requires covariate data to be available on a geography compatible with MSOAs. A range of data sources were used in the modelling process that included variables considered to be related to the propensity of a household having income below a threshold. They are:

- Census, 2011
- Department for Work and Pensions benefit claimant counts, August 2011
- Valuation Office Agency Council Tax Bandings, 27 March 2011
- Her Majesty's Revenue and Customs, Child Tax Credit and Working Tax Credit, 2011
- Communities and Local Government, Change of ownership by dwelling price, 2009
- Regional/country identification variable

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

The data used are as close to the reference time period of the target income estimates as possible (i.e. for 2011/12). Administrative data are collected primarily for government administrative processes and may change over time. The DWP data sources for benefit claimants and HMRC data sources for Tax Credits have changed since the reference time period of these estimates. More information about the variables considered for inclusion in the model and the recent changes to the sources is provided in Appendix B.

3.3 The statistical model

Binary response 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 as the response variable households in poverty (1 if the household is in poverty and 0 otherwise) and the area level covariates as explanatory variables. The models relate the survey variable of interest (measured at household level) to the covariates that correspond to the small area in which the household is located. Once fitted, the models can be used to produce estimates of the target variables at the small area level and their confidence intervals, i.e. the models can be used to produce MSOA-level estimates of:

- the proportion of households in poverty AHC (per cent of households with net equivalised income AHC below 60 per cent of the national median income);
- the proportion of households in poverty BHC (per cent of households with net equivalised income BHC below 60 per cent of the national median income).

The SAEP methodology has been developed to produce model based estimates of continuous or binary variables contained in social surveys. The mean income model based estimates are produced using a linear model taking the continuous survey variable weekly household income (logarithm transformed) as the response variable.

In order to model the distribution of low income within MSOAs the income data at the household level is transformed into a binary variable; 1 if the household income is below 60 per cent of the national median and 0 otherwise. A binomial model using the logit link function that takes into account that each household belongs to a specific area can be developed to relate the binary variable to area level covariates.

Note that the sampling area in the survey is the PSU but the estimation area is the MSOA. As the FRS uses a clustered sample design, PSUs and MSOAs can cross cut each other. This means that the area level variation in the model has to be measured using the PSU (as this is the area where the data come from). The model assumes that variation for MSOAs is similar to variation for PSUs as PSUs and MSOAs are of similar size in terms of households. This allows us to use the variance associated with PSUs in error calculations relating to MSOAs. This assumption was assessed (SAEP, 2003) using LFS data (not clustered) which showed that within PSU variability was similar to within ward variability (MSOAs now replace wards but wards and MSOAs are similar).

A household level model is fitted because the sampling area is different to the estimation area and within each MSOA PSUs can overlap. This means that covariate information (at MSOA area level) is available for households and can vary for those in same PSU. The underlying model is a two level model given by:

$$y_{id} \sim Binomial(1, \pi_{id})$$

$$y_{id} = \pi_{id} + e_{id}$$

$$logit(\pi_{id}) = \alpha + \boldsymbol{X}_d^T \boldsymbol{\beta} + u_j$$

where

y_{id} is the survey variable for household *i* in MSOA *d*, so is the poverty indicator for household *i* in MSOA *d*;

j is the sampling area (PSU);

π_{id} is the expected probability of household *i* in MSOA *d* being in poverty (i.e. having a poverty indicator of 1) ;

$\mathbf{X}_d$ is a vector of values for area *d* of a set of covariates;

u_j is the area level residual for primary sampling unit *j* (sampling area); assumed to have expectation 0 and variance σ_u^2 ;

and e_{id} is within area residual for household *i* in MSOA *d* with expectation 0 and variance σ_e^2 .

Models are fitted on the sample data and using covariates in areas for which a sample is present. However, as covariates are available for all areas, a synthetic estimator of the proportion of households with an income level below 60 per cent of the national median can be produced for all areas from the fitted model. This is given by

$\hat{\pi}_d^{synth} = \dfrac{exp\left(\hat{\alpha} + \boldsymbol{X}_d^T \hat{\boldsymbol{\beta}}\right)}{1 + exp\left(\hat{\alpha} + \boldsymbol{X}_d^T \hat{\boldsymbol{\beta}}\right)}$, where $\hat{\alpha}$ is the estimate of α and $\hat{\boldsymbol{\beta}}$ is the vector of estimated coefficients for the covariates.

Note, a composite estimator using the area level residuals, u_j, cannot be obtained because these residuals are associated with the PSU and not the MSOA. The 95 per cent confidence interval for the an area prediction in the logit scale is given by

$$\left(\hat{\alpha} + \mathbf{X}_d^T \hat{\boldsymbol{\beta}}\right) \pm 1.96 \sqrt{\left(\hat{\sigma}_u^2 + Var\left((1, \mathbf{X}_D)^T \begin{pmatrix} \hat{\alpha} \\ \hat{\boldsymbol{\beta}} \end{pmatrix}\right)\right)}$$

Under the SAEP methodology the area level variance, σ_u^2 , is added to the standard error term to provide an approximation to the 95 per cent confidence limits. The anti logit of these limits are taken to produce a 95 per cent confidence interval for $\hat{\pi}_d^{synth}$.

4.Developing the models

4.1 Introduction

The previous section of the report introduced the statistical model and datasets used to produce model based estimates of proportion of households below 60% of median income. This section describes the model fitting procedures. The controlled version of the variable selection approach is described in Section 4.2 followed by a description of the chosen AHC and BHC models for 2011/12 in Section 5.

Before proceeding it is important to highlight that explanatory analysis indicated that no unique covariate in the available auxiliary datasets showed a high correlation with the response variable (correlations between logit of the proportion of household below threshold and the covariates were calculated), whereas the covariates are highly correlated to each other. This is the worst scenario an analyst could encounter, a database full of covariates, all conveying the same information about the response variable, but none of them presenting a well defined relationship with the variable of interest.

4.2 Model Selection Procedure

The controlled model selection procedure described here was carried out separately for income AHC and BHC. Alongside the controlled selection, the automatic model selection was also run and the model considered.
The full set of covariates was first separated into two groups: untransformed covariates and logit transformed covariates. Within these two groups, covariates from different datasets (e.g. census, DWP, HMRC, etc.) were run through an automatic model selection separately. This was carried out with a single level model. In order to choose the final model, the variables identified separately from each data source (untransformed and/or transformed) as potential good predictors for the poverty indicator were jointly included in a model and a controlled backwards elimination was carried out. If both an untransformed and transformed version of one covariate were selected then the less significant of the two would be removed from the model.
The final selection of covariates from the backwards model selection was used to create possible interactions and these interactions were also tested in a stepwise selection. The covariates and interactions were fitted in a multilevel model for the binary response and non-significant variables were removed, except when they were involved in a significant interaction.
The multilevel model was fitted with postcode sectors at the higher level and households at the lower level, as outlined in Section 3.3. This and the single level models used for the model selection were fitted using the statistical software SAS. Country/regional indicator terms are forced into the model (whether significant or not) as an attempt to control for region differences and to reduce the amount of calibration that would be necessary for benchmarking the model based estimates to the published HBAI estimates at country/region level.

In previous applications of the poverty modelling the census variables were grouped to highlight how they convey information about the similar characteristics of the areas. Although the variables were not the same for every model, all models included (before the backwards model selection) at least one Census covariate representing the same underlying dimensions/area characteristics: socio-economic classification or social grade; nationality/ethnicity of the resident population; labour market status; dwelling characteristics and the proportion of population in working age. The conclusion drawn was that although the models may look different, they present a sort of stability in relation to which information is conveyed by the selected Census covariates.

For 2011/12, the inclusion of these census variables groupings was also considered to see if it would improve the model. The AHC model had not selected any labour market status census variables to go into the backwards selection, so they were all included and one of the variables (pretired) entered the final model.

The measure of model adequacy used to compare competing models was the percentage of between area variability explained by the covariates in the model calculated as:

% between area variability explained = $\left(1-\frac{\sigma_u^2 \text{ (full model)}}{\sigma_u^2 \text{ (null model)}}\right)\times 100$.

In addition, the percentage of between area variability explained by one covariate in the presence of all covariates was calculated as:

$$\left(1-\frac{\sigma_u^2 \text{ (full model)}}{\sigma_u^2 \text{ (model excluding 1 covariate)}}\right)\times 100$$

The chosen models and corresponding adequacy measures are presented in Section 5.

5. The fitted model

5.1 AHC Model

The final model for the proportion of households below income threshold (AHC) for 2011/12 is given by:

$logit(\hat{\pi}_{id})$ =

- 1.337(0.088) Constant
+ 0.243(0.137) northest
+ 0.002(0.115) northwst
+ 0.007(0.118) york
+ 0.145(0.119) eastmid
- 0.030(0.112) westmid
+ 0.007(0.114) east
+ 0.055(0.144) wales
+ 0.053(0.106) southest
+ 0.067(0.122) southwst

Region/Country

- 1.255(0.377) pgroupab
- 4.817(0.939) pcouple
- 4.238(1.056) pintocc
+ 0.601(0.187) pnonwbri
+ 4.268(0.983) pretired
+ 0.763(0.231) lnphrpmale
+ 0.285(0.094) lnisptotal
- 0.252(0.103) lndlamah
- 23.684(9.96) pintocc_pcouple
+ 0.395(0.147) lnisptotal_wales

The figures in parentheses are the standard errors (s.e) of the estimated coefficients. Table 1 contains a key to the labels of the covariates. The test statistic is defined as T-ratio = $\hat{\beta} / \text{s.e}$.

Table 1 Key to covariates included in the model for households in poverty AHC, 2011/12

Covariate Name	Label	Source	T-ratio
northest	North East	Country/regional indicators	1.77
northwst	North West	Country/regional indicators	0.01
york	Yorkshire and The Humber	Country/regional indicators	0.06
eastmid	East Midlands	Country/regional indicators	1.22
westmid	West Midlands	Country/regional indicators	-0.27
east	East of England	Country/regional indicators	0.06
wales	Wales	Country/regional indicators	0.38
southest	South East	Country/regional indicators	0.49
southwst	South West	Country/regional indicators	0.55
pgroupab	Proportion of people aged 16 to 74 whose approximated social grade is AB	Census	-3.33
pcouple	Proportion of people in households that are living in a couple	Census	-5.13
pintocc	Proportion of people aged 16 to 74 whose NS-SEC is 'intermediate'	Census	-4.01
pnonwbri	Proportion of people who are 'Not White British'	Census	3.21
pretired	Proportion of people aged 16 to 74 who are retired	Census	4.34
lnphrpmale	Logit of Proportion of household reference persons who are male	Census	3.30
lnisptotal	Logit of Proportion of people aged 16 and over claiming Income Support	DWP	3.04
lndlamah	Logit of Proportion of people claiming Disability Living Allowance: Mobility Award Higher	DWP	-2.44
pintocc_pcouple	Interaction between pintocc and pcouple	Interaction Term	-2.38
lnisptotal_wales	Interaction between lnisptotal and wales	Interaction Term	2.69

Note: London is arbitrarily chosen as the baseline for the region indicators

With no covariates included in the model the estimated standard residual area variance $\hat{\sigma}_u^2$ and standard error is 0.2366 (0.0286) compared with 0.0418 (0.0209) obtained when the significant covariates are included in the model. Therefore, these covariates together account for 82.3 per cent of the total between area variance. Although the AHC model has performed less well compared with the both the BHC and previous AHC models, the reduction in between area variability explained by the chosen model is not sufficiently large to cause concern about the overall quality of the model.

To understand the decomposition of the between area variance, the model can be fitted by including each covariate on its own. The covariates that individually account for most of the between area variability are pcouple (Census) and lnisptotal (DWP) which each account for 59 per cent and 55 per cent respectively of the between area variability.

Table 2 below shows how much of the between area variability is explained by each covariate in the absence of all other covariates.

Table 2 Decomposition of area level variance AHC, 2011/12

Covariate Name	% between area variability explained by covariate on its own
Region	8.71
pgroupab	26.18
pretired	24.89
pcouple	59.19
pintocc	35.20
pnonwbri	35.44
lnphrpmale	25.64
lnisptotal	54.69
lndlamah	6.75

Source: Office for National Statistics

The covariates with the largest percentage of between area variability explained are pcouple, lnisptotal and pnonwbri. Pcouple is negatively correlated with estimated poverty proportion with lnisptotal and pnonwbri both being positively correlated.

5.2 BHC Model

The final model for the proportion of households below income threshold (BHC) for 2011/12 is given by:

$logit(\hat{\pi}_{id})$ =

- 1.810(0.092) Constant
+ 0.314(0.145) northest
+ 0.209(0.116) northwst
+ 0.270(0.118) york
+ 0.292(0.121) eastmid
+ 0.271(0.113) westmid
+ 0.252(0.117) east
+ 0.382(0.138) wales
+ 0.270(0.113) southest
+ 0.272(0.128) southwst

Region/Country

+ 0.491(0.157) pnonwbri
+ 0.052(0.024) ewtrns
+ 0.325(0.049) lnpgroupde
- 0.504(0.098) lnpemployd
+ 1.212(0.383) pnonwbri_york

The figures in parentheses are the standard errors (s.e) of the estimated coefficients.
Table 3 contains a key to the labels of the covariates. The test statistic is defined as T-ratio = $\hat{\beta}/\mathrm{s.e}$.

Table 3 Key to covariates included in the model for households in poverty BHC, 2011/12

Covariate Name	Label	Source	T-ratio
northest	North East	Country/regional indicators	2.17
northwst	North West	Country/regional indicators	1.80
York	Yorkshire and The Humber	Country/regional indicators	2.28
eastmid	East Midlands	Country/regional indicators	2.42
westmid	West Midlands	Country/regional indicators	2.39
East	East of England	Country/regional indicators	2.15
Wales	Wales	Country/regional indicators	2.77
southest	South East	Country/regional indicators	2.39
southwst	South West	Country/regional indicators	2.13
pnonwbri	Proportion of people who are 'Not White British'	Census	3.14
Ewtrns	Transactions by Dwelling Type; Total Sales	DCLG	2.10
lnpgroupde	Proportion of people aged 16 to 74 whose approximated social grade is D	Census	6.63
lnpemployd	Logit of Proportion of people aged 16 to 74 who are employed or self-employed	Census	-5.14
pnonwbri_york	Interaction between pnonwbri and york	Interaction term	3.16

Note: London is arbitrarily chosen as the baseline for the region indicators

With no covariates included in the model the estimated standard residual area variance $\hat{\sigma}_u^2$ and standard error is 0.1176 (0.0260) compared with 0.0069 (0.0222) obtained when the significant covariates are included in the model. Therefore these covariates together account for 94.2 per cent of the total between area variance.

To understand the decomposition of the between area variance, the model can be fitted by including each covariate on its own. The covariates that individually account for most of the between area variability are lnpemployd (Census) and lnpgroupde (Census) which each account for 69 per cent and 68 per cent respectively of the between area variability.

Table 4 shows how much of the between area variability is explained by each covariate in the absence of all other covariates.

Table 4 Decomposition of area level variance BHC, 2011/12

Covariate Name	% between area variability explained by covariate on its own
Region	6.77
pnonwbri	11.07
ewtrns	6.19
lnpgroupde	68.01
lnpemployd	69.31

The covariates with the largest percentage of between area variability explained are lnpemployd and lnpgroupde. Lnpemployd relates to people aged 16 to 74 who are employed or self-employed and as expected this is negatively correlated to the estimated proportion of poverty. Lnpgroupde relates to people aged 16 to 74 whose approximated social grade is D (Semi-skilled and unskilled manual occupations; unemployed and lowest grade occupations) and this is positively correlated with estimated poverty proportion.

6. Model estimates

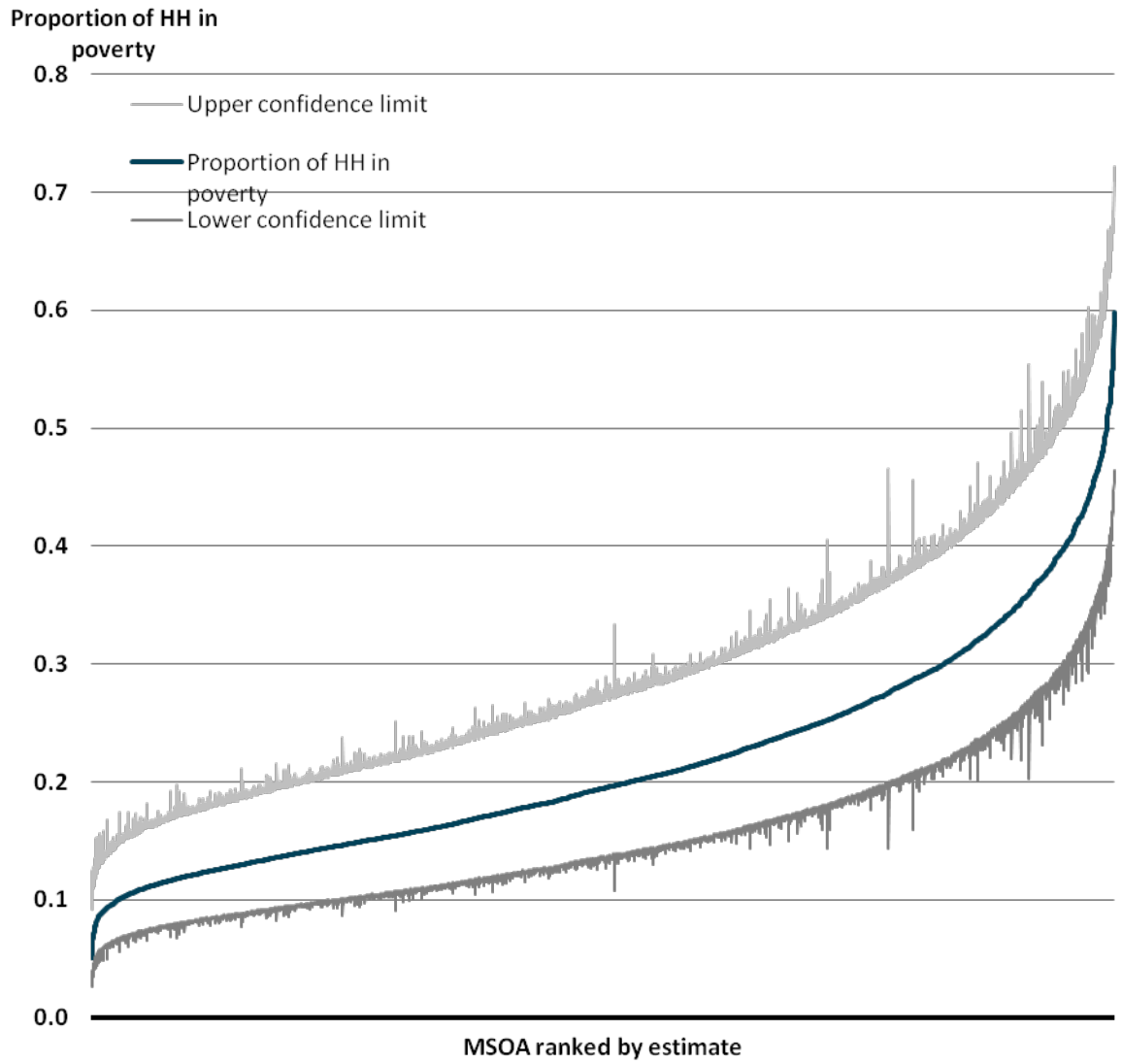
6.1 AHC Estimates

The model described in Section 5.1 was used to produce the model based estimates of the proportion of households below an income AHC threshold for all MSOAs in England and Wales. Figure 1 below presents the ranked estimates (calibrated to survey region and Wales estimates[1]) that are also displayed in a map available in Appendix A. The proportion in poverty (blue line) is shown together with the top and bottom of the confidence intervals for the MSOAs.

The upper and lower confidence limits show considerable variation indicating that the estimates vary greatly. The gradient of the estimate line is steep for most of its range. This means that MSOAs at the top and bottom of the distribution can be statistically distinguished with 20.8 per cent (1,495) of MSOAs at the lowest ranks being statistically distinguishable from the same number of the highest ranks (for example their confidence intervals do not overlap).

[1] Calibration ratios are presented in Section 7.

Figure 1 **Estimates and confidence intervals of the proportion of households in each MSOA below 60 per cent median AHC, 2011/2012**



Source: Office for National Statistics

6.2 Precision of AHC estimates

Figure 2 shows the distribution of the coefficients of variation for 2011/12. Ideally the coefficients of variation (CVs) should be below 20 per cent for estimates to be considered precise. For the vast majority of areas (6383 MSOAs) the CVs were below 20 per cent. Eighteen MSOAs had coefficients of variation over 30 per cent, all of which are in Wales.

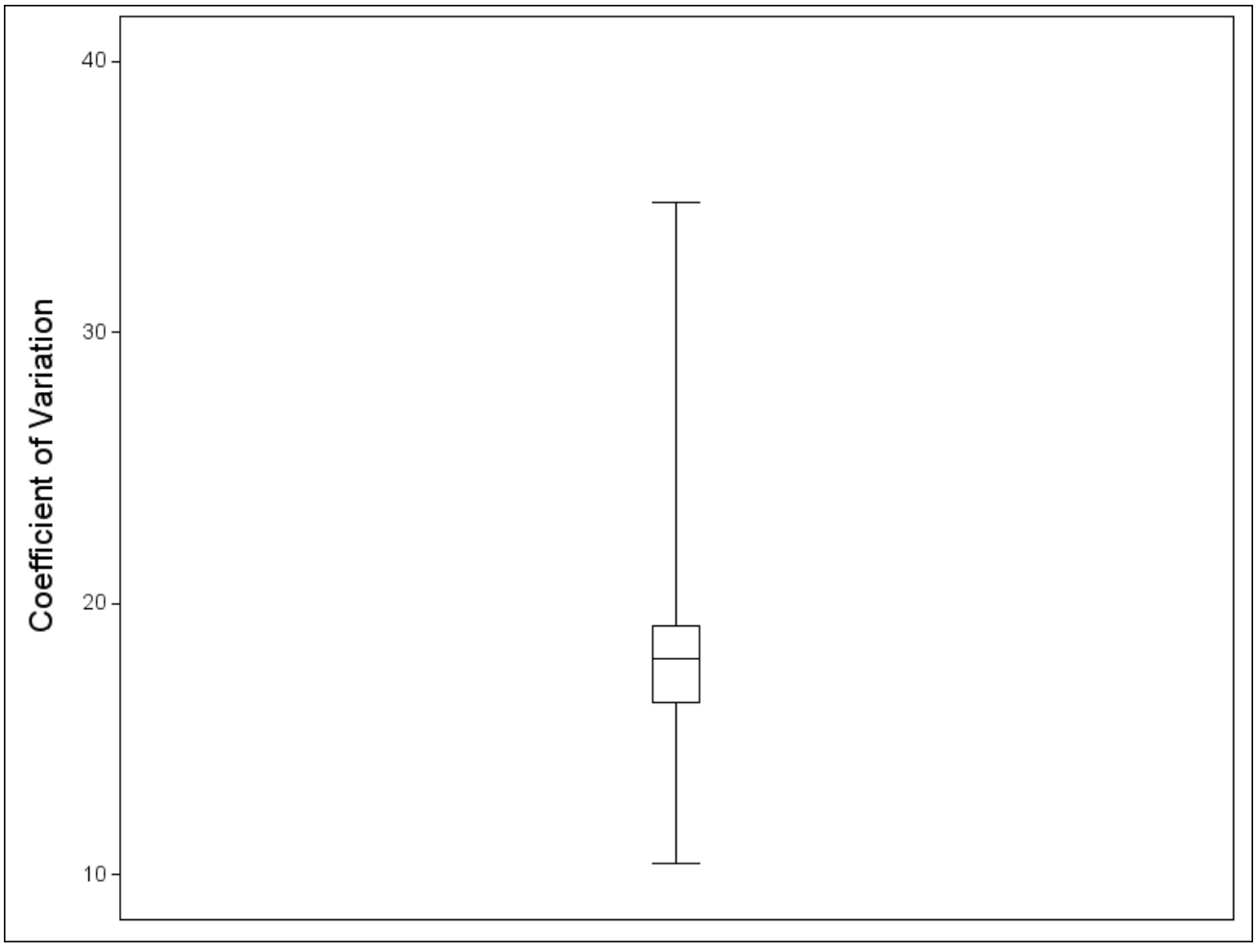
These CVs are somewhat larger than the BHC model and the previous AHC model, indicating that the fit of the AHC model is not quite as good as either the BHC model or the previous AHC model. However, the increase in the number of high CVs with the chosen model is not sufficient to cause concern about the overall quality of the model. Table 5 summarizes the distribution of the coefficients of variation.

Table 5 Summary of coefficients of variation (AHC), 2011/2012

Minimum	Lower quartile	Median	Upper quartile	Maximum
10.41	16.36	17.95	19.17	34.81

Source: Office for National Statistics

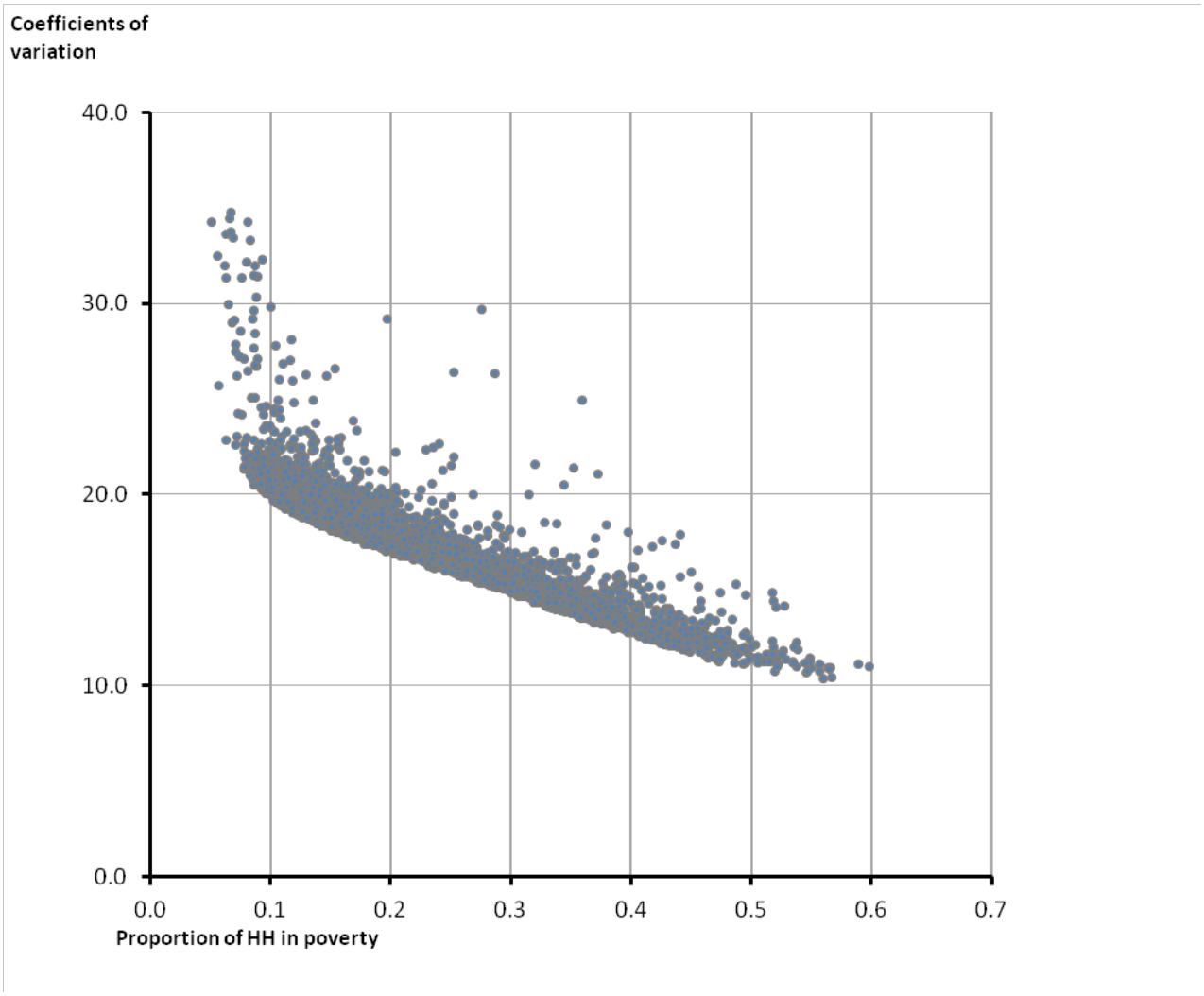
Figure 2 Distribution of coefficients of variation (AHC), 2011/2012



Source: Office for National Statistics

Figure 3 displays the coefficients of variation against the modelled estimates of the proportion of households in poverty. The more precise estimates are those where higher proportions of households in poverty are predicted.

Figure 3 Coefficients of variation against modelled estimates of proportion of households in poverty (AHC), 2011/2012



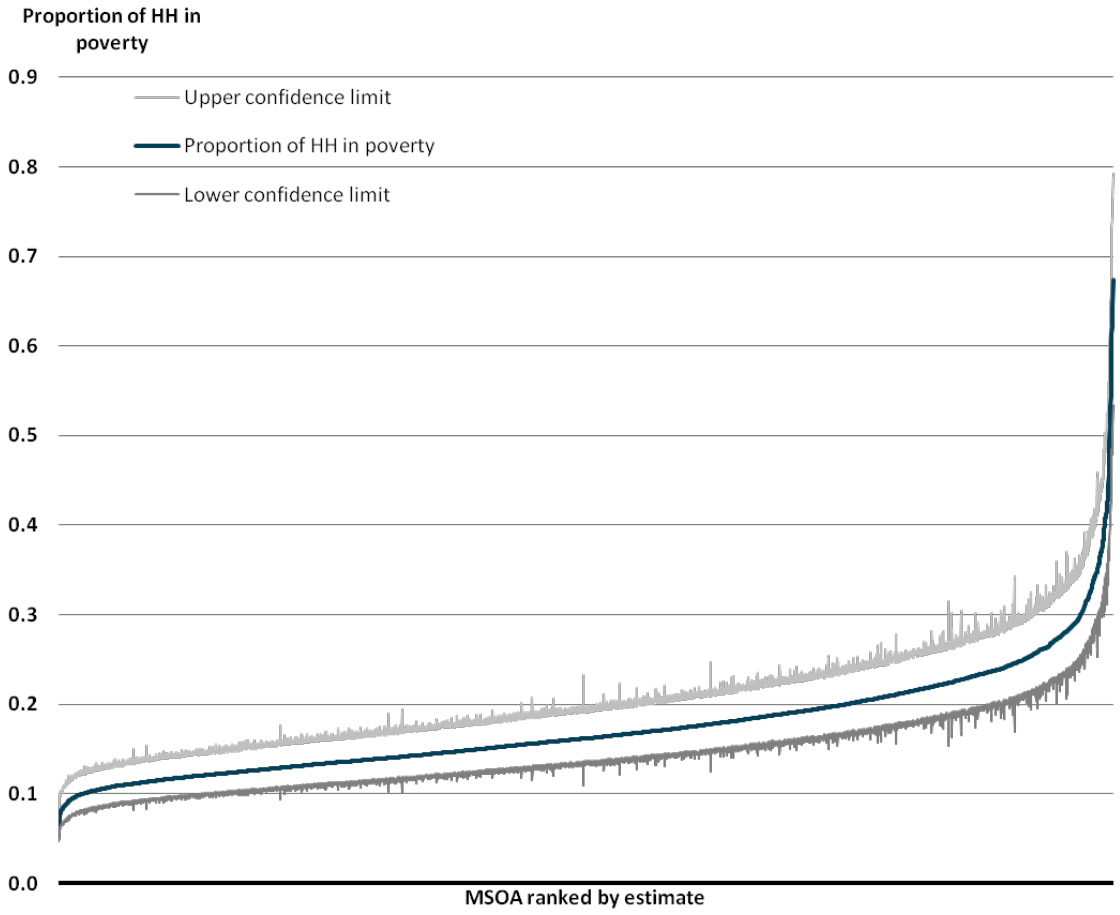
Source: Office for National Statistics

6.3 BHC Estimates

The model described in Section 5.2 was used to produce the model based estimates of the proportion of households below an income BHC threshold for all MSOAs in England and Wales. Figure 4 below presents the ranked estimates (calibrated to survey region and Wales estimates[2]) that are also displayed in a map available in Appendix A. The proportion in poverty (blue line) is shown together with the top and bottom of the confidence intervals for the MSOAs.

The upper and lower confidence limits show considerable variation indicating that the estimates vary greatly. The gradient of the estimate line is steep for most of its range. This means that MSOAs at the top and bottom of the distribution can be statistically distinguished with 26.8 per cent (1,927) of MSOAs at the lowest ranks being statistically distinguishable from the same number of the highest ranks (for example their confidence intervals do not overlap).

Figure 4 **Estimates and confidence intervals of the proportion of households in each MSOA below 60 per cent median BHC, 2011/2012**



Source: Office for National Statistics

[2] Calibration ratios are presented in Section 7.

6.4 Precision of BHC estimates

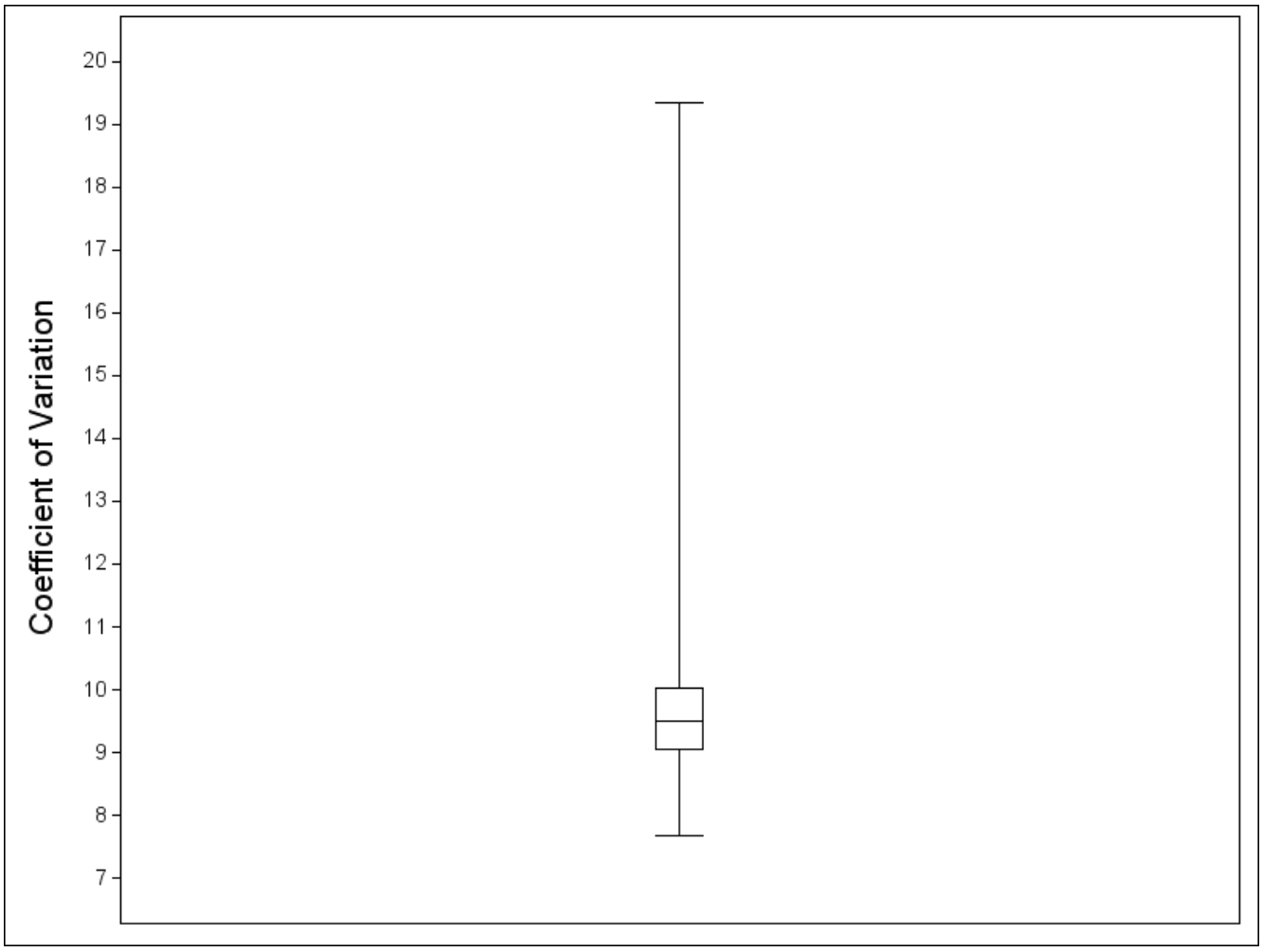
Figure 5 shows the distribution of the coefficients of variation for 2011/12. Ideally the coefficients of variation should be below 20 per cent for estimates to be precise enough. For all of the areas (7,201 MSOAs) the coefficients of variation are below 20 per cent. Table 6 summarizes the distribution of the coefficients of variation.

Table 6 **Summary of coefficients of variation (BHC), 2011/2012**

Minimum	Lower quartile	Median	Upper quartile	Maximum
7.69	9.05	9.49	10.04	19.35

Source: Office for National Statistics

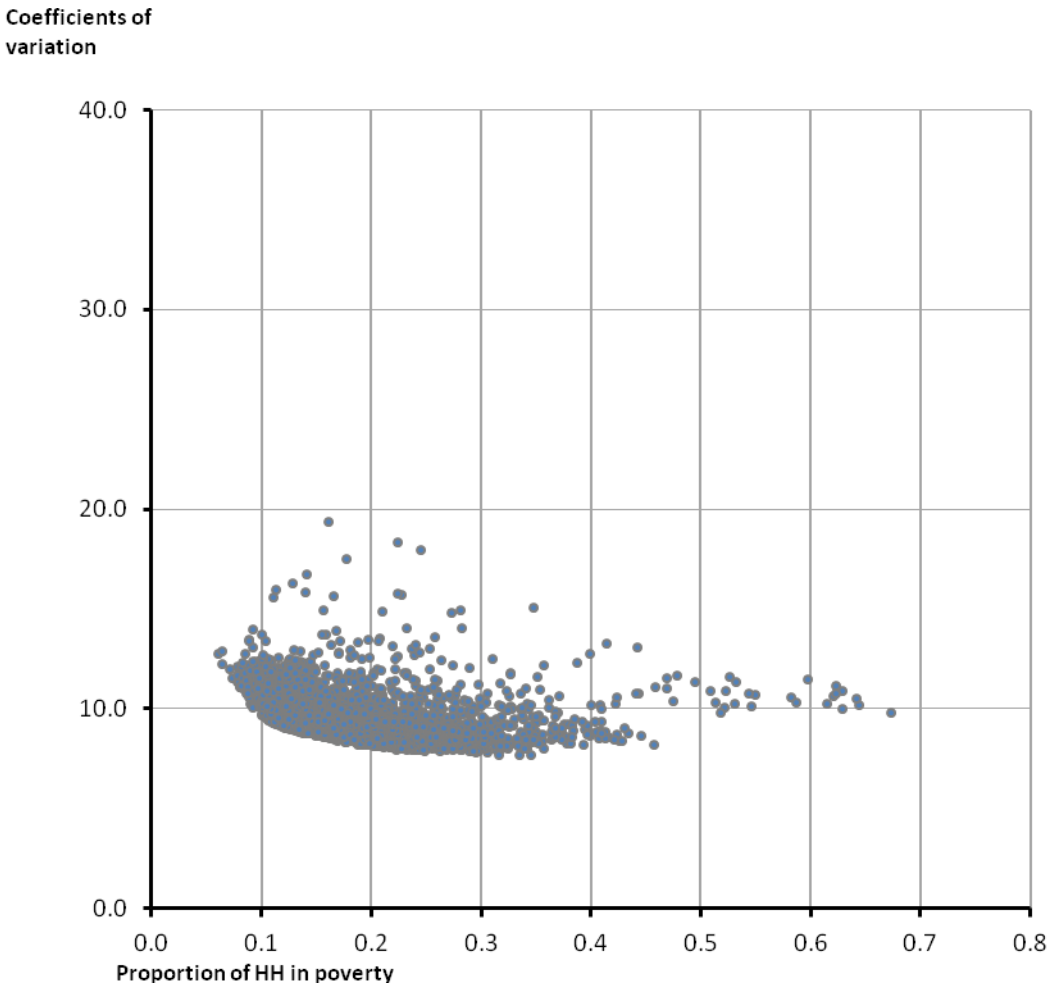
Figure 5 **Distribution of coefficients of variation (BHC), 2011/2012**



Source: Office for National Statistics

Figure 6 displays the coefficients of variation against the modelled estimates of the proportion of households in poverty. Unlike the AHC estimates, there is not much difference in the precision of the lower and higher BHC estimates of the proportion of households in poverty.

Figure 6 **Coefficients of variation against modelled estimates of proportion of households in poverty (BHC), 2011/2012**



Source: Office for National Statistics

7. Model diagnostics

This section describes the different diagnostic checks that have been used to assess the appropriateness of the models developed. The diagnostic checks employed here are those developed by the ONS for small area estimation. Each diagnostic test is described and the results displayed for 2011/12 (AHC & BHC) for England and Wales.

7.1 Residual vs. Model Estimates Diagnostic Plot

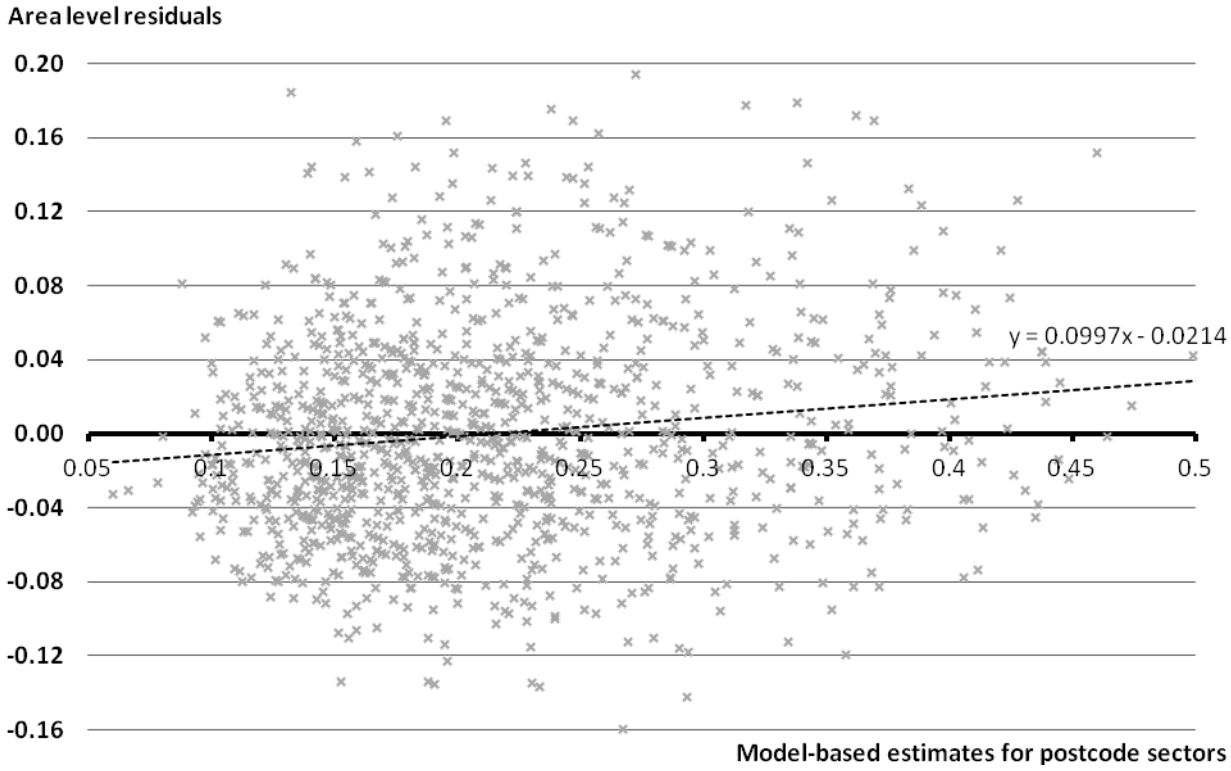
A plot of model estimates against model residuals at the MSOA level is a method of checking that the model assumptions are satisfied and the model accurately describes the population. Here we are testing for two things: model mis-specification and non-constant variance of the residuals (heteroscedasticity). If any pattern remains in the residuals this implies model mis-specification e.g. a covariate influential to predicting the response variable 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.

Due to the structure of the models, area level residuals refer to postcode sectors (PCS). For the plot of area level residuals we require model based estimates at the PCS level, however, covariates are by MSOA and not PCS. In order to form model based estimates for PCS for the plot an approximate method is used. The estimates of poverty proportion are aggregated to the PCS level. For this residual diagnostic we are making the assumption that the results at PCS level would highlight any problems at MSOA-level.

7.1.1 AHC estimates

Figure 7 below shows that there is a slight pattern to the AHC residuals, as the intercept and slope of the regression line are both significantly different from zero. Therefore, there isn't evidence in favour of the constant variance assumption after modelling at the area level. However, results from the other diagnostics support the AHC model.

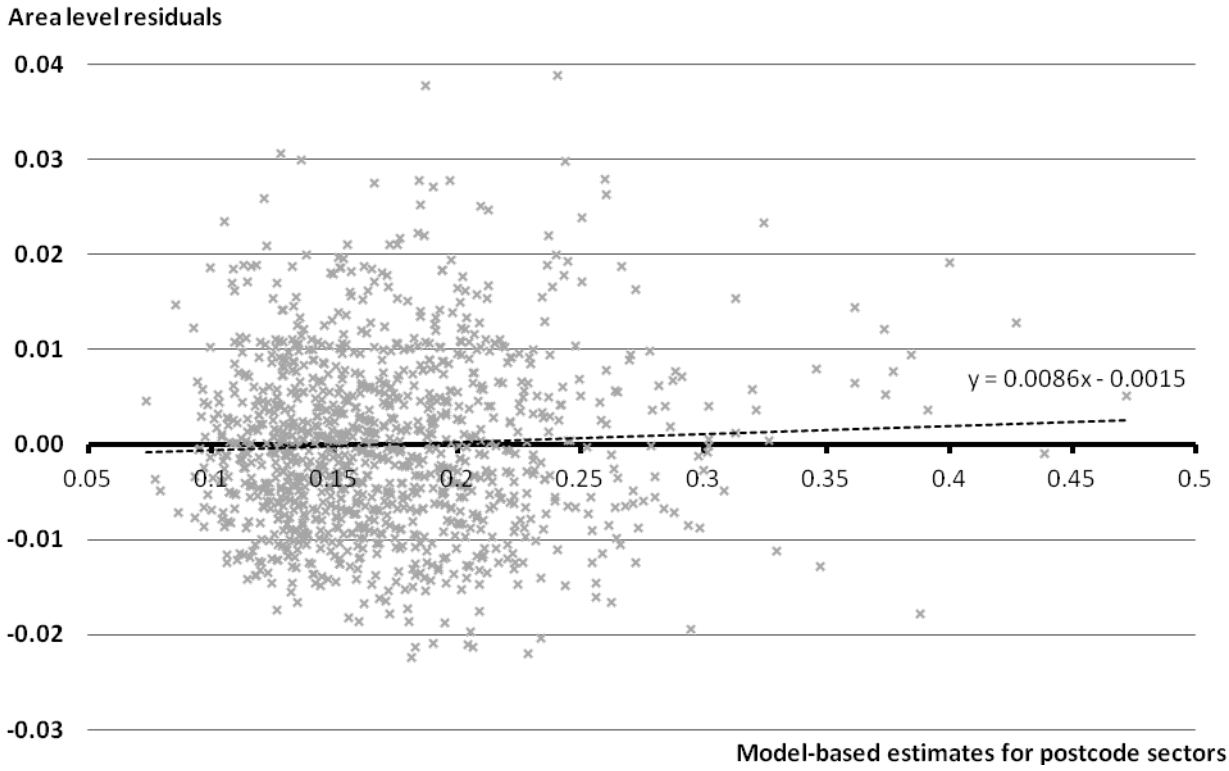
Figure 7 **Area level residuals, AHC 2011/2012**



7.1.2 BHC estimates

Figure 8 below shows that there is no pattern to the BHC residuals, as the intercept and slope of the regression line are not significantly different from zero. Therefore there is evidence in favour of the constant variance assumption after modelling at the area level.

Figure 8 Area level residuals, BHC 2011/2012



Source: Office for National Statistics

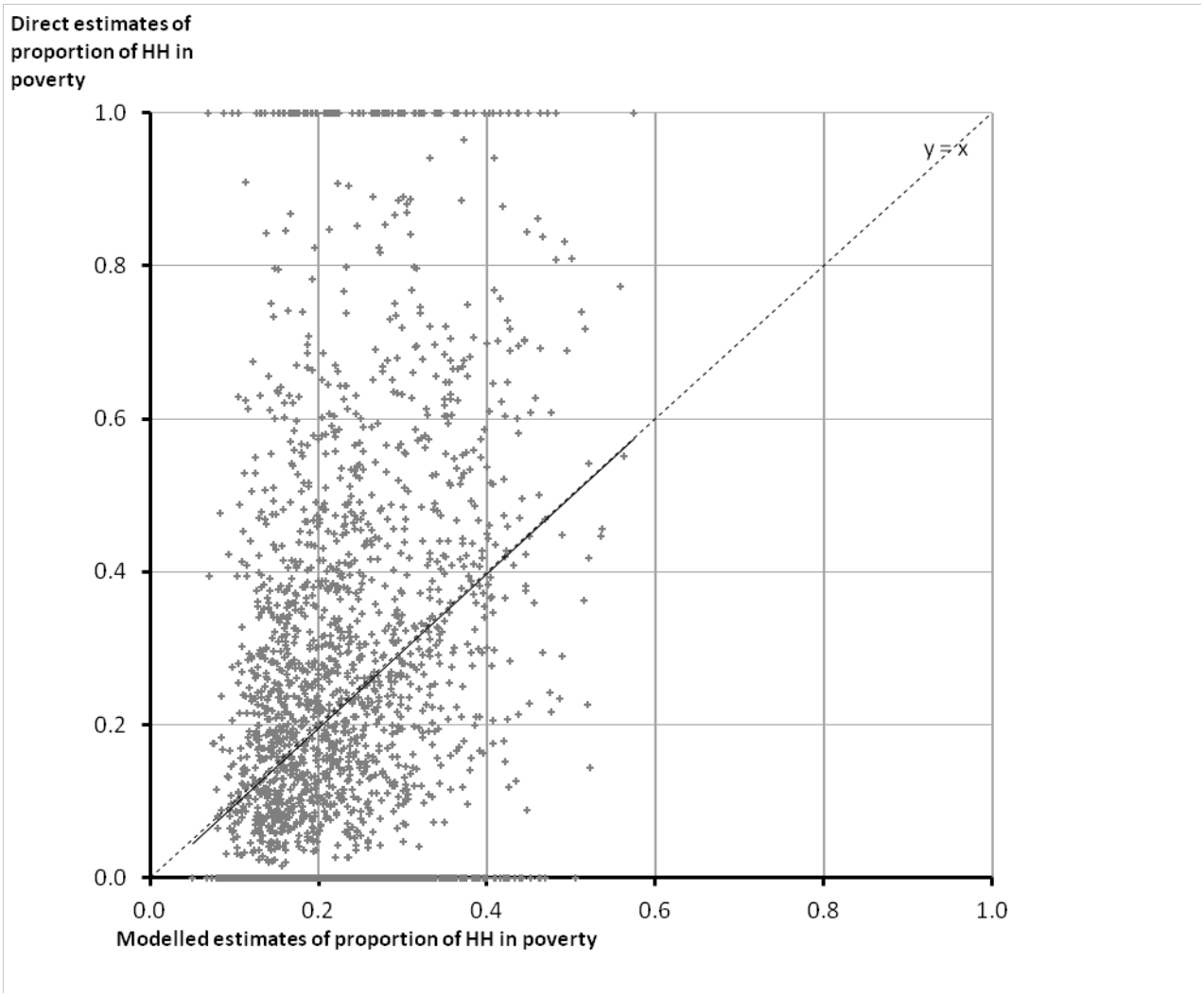
7.2 Model vs. Sample Estimates Diagnostic Plot

A plot of direct survey estimates (y-axis) against model-based estimates (x-axis), for MSOAs for which there is a sample, is one method of assessing whether the relationship between the target variable and the covariates has been specified properly. For good model-based estimates, the direct estimates will be randomly distributed around the estimates and the regression line between the two will be very close to the line y=x. If the relationship between the target variable and the covariates has been mis-specified or mis-estimated 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.

7.2.1 AHC estimates

Figure 9 below displays the plot of direct survey estimates (AHC) against AHC model based estimates and also the y=x (dashed) and fitted lines.

Figure 9 **Direct estimates against modelled estimates (AHC), 2011/2012**

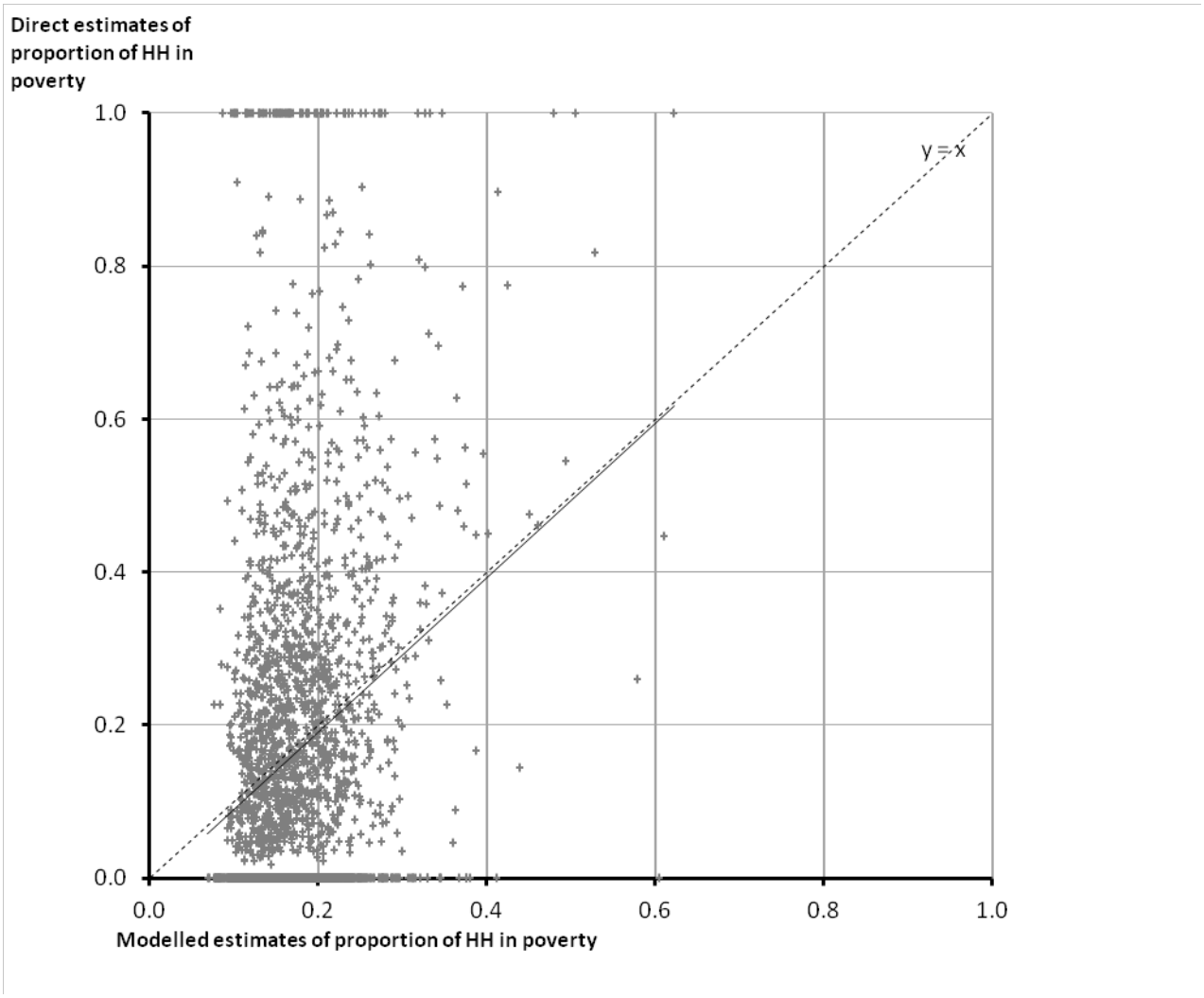


It is clear that there is much wider variation for the direct survey estimates than for the model based estimates. This is due to the fact that the sample MSOAs have an extremely small sample size. This results in 37 per cent of the direct estimates being equal to zero and 4 per cent having a value of one. A test of significance indicates that the regression slope is not significantly different from one and that the constant is not significantly different from zero. This indicates that the modelled estimates are similar to the direct estimates and that no bias is present in the modelled estimates.

7.2.2 BHC estimates

Figure 10 below displays the plot of direct survey estimates (BHC) against BHC model based estimates and also the y=x (dashed) and fitted lines.

Figure 10 **Direct estimates against modelled estimates (BHC), 2011/2012**



Source: Office for National Statistics and Family Resources Survey, Department for Work and Pensions

As for AHC, there is much wider variation in the direct survey estimates than for the model based estimates. In this case, 43 per cent of the direct estimates are equal to zero and 3 per cent having a value of one. A test of significance indicates that the regression slope is not significantly different from one and that the constant is not significantly different from zero. This indicates that the modelled estimates are similar to the direct estimates and that no bias is present in the modelled estimates.

7.3 Coverage Diagnostic

The purpose of this diagnostic is to examine the validity of the confidence intervals for the model-based estimates. For those MSOAs in sample, there will be direct survey estimates with associated 95% confidence intervals. However in view of the binary nature of the variable, MSOAs with direct estimates lying close to either extreme or with very few respondents will cause unreliable direct standard errors. Therefore for this diagnostic and for the following Wald statistic, MSOAs with less than five respondents or where direct estimates of poverty are under 7.5% or over 92.5% are excluded.

The diagnostic measures the overlap between the direct confidence intervals and the corresponding model-based estimate confidence intervals, i.e. it measures the percentage of MSOAs for which the model and direct confidence intervals overlap.

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

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

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

Consequently, this diagnostic takes $z(\alpha) = 1.96$, calculates $z(\beta)$ using the above formula, with σ_X replaced by the estimated standard error of the model-based estimate and σ_Y replaced by the estimated standard error of the direct estimate and then computes the overlap proportion between the corresponding $z(\beta)$ -based confidence intervals. For $z(\alpha) = 1.96$ this proportion should be 95%. Any significant deviation from a 95% overlap will indicate that the model based confidence intervals are generally too wide or too narrow.

The analysis shows that an overlap occurs in 1014 [890] out of 1040 [918] MSOAs (which is also greater than the required 95%). A pooled variance has been used to calculate the confidence intervals for the direct estimates (see Appendix F) and this will result in an overestimation of these confidence intervals and hence a coverage percentage slightly greater than 95% is not a surprising result.

7.4 Wald Statistic

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

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

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

The goodness-of-fit statistic for the model developed here is 1065.1 [822.3] on 1040 [918] degrees of freedom, this has a p-value of 0.29 [0.95]. There is no significant evidence to reject a χ^2 distribution. Therefore, for both AHC and BHC poverty, there is no significant difference between the expected values of the model-based estimates and the direct survey estimates.

7.5 Stability Analysis

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

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

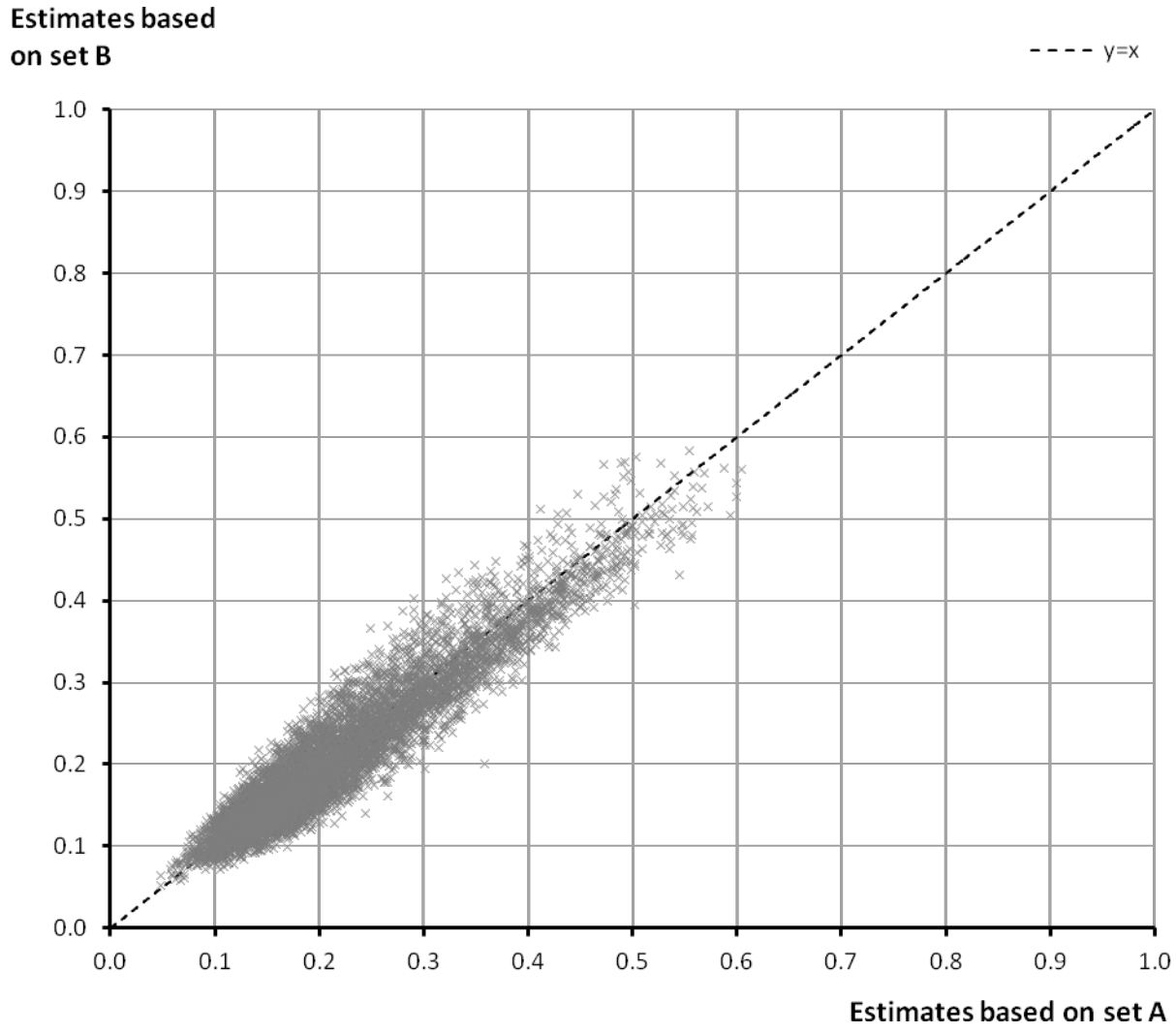
$$\text{RRMSE} = \sqrt{\sum_i \frac{1}{n}\left(\frac{\hat{Y}_B - \hat{Y}_A}{\hat{Y}_A}\right)^2}$$

where $\hat{Y}_A$ and $\hat{Y}_B$ are the model-based estimates calculated using regression coefficients $\hat{\beta}_{k_A}$ and $\hat{\beta}_{k_B}$ respectively and n is the total number of MSOAs.

7.5.1 AHC estimates

The median RRMSE for the 10 repetitions for 2011/12 is 0.023. The RRMSE shows that the two sets of estimates are fairly similar and that there is stability in the model. A RRMSE greater than 0.5 is considered as an indication of instability. Figure 11 illustrates a comparison of the AHC model based estimates obtained with the y=x line, for one set of estimates.

Figure 11 **Comparison of two sets of estimates for stability analysis, (AHC) 2011/2012**

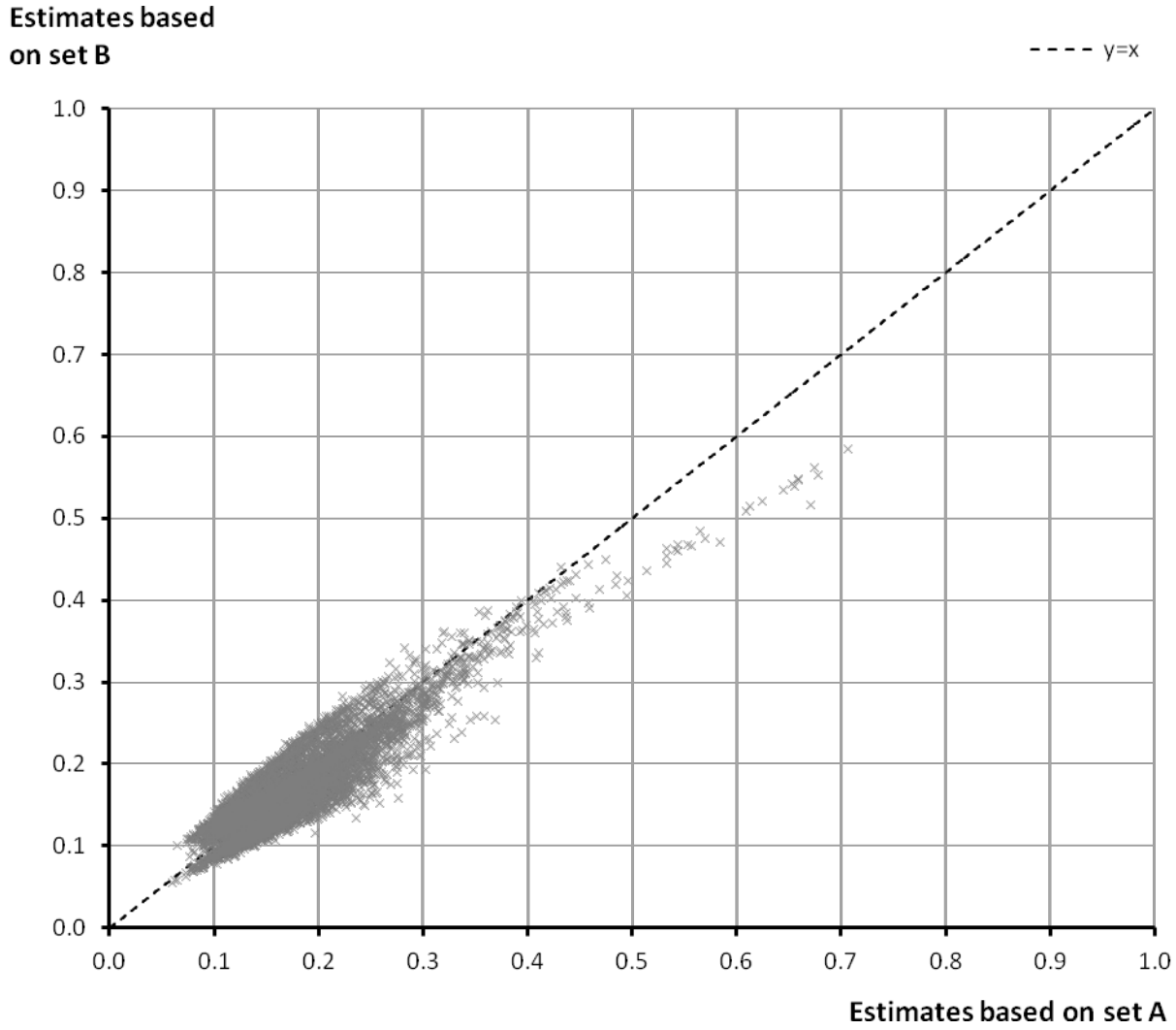


Source: Office for National Statistics

7.5.2 BHC estimates

The median RRMSE for the 10 repetitions for 2011/12 is 0.017. The RRMSE shows that the two sets of estimates are fairly similar and that there is stability in the model. A RRMSE greater than 0.5 is considered here as an indication of instability. Figure 12 illustrates a comparison of the BHC model based estimates obtained with the y=x line, for one set of estimates.

Figure 12 **Comparison of two sets of estimates for stability analysis, (BHC) 2011/2012**



7.6 Benchmarking

The estimates for the AHC and BHC models are calibrated using the survey region and Wales totals and model totals. Tables 7 and 8 show the survey and model totals and the calibration ratios used to adjust the model estimates and their confidence intervals. Calibration ratios are above and below one indicating that there is no bias in the modelled estimates.

Table 7 Benchmarking results for AHC model, 2011/2012

Country/GOR	Number of MSOAs	Aggregated survey total	Model total	Ratio of survey to model total
North East	340	0.223	0.244	0.915
North West	924	0.213	0.212	1.004
Yorkshire	692	0.215	0.209	1.026
East Midlands	573	0.224	0.224	0.997
West Midlands	735	0.217	0.218	0.996
East	736	0.186	0.187	0.998
London	983	0.277	0.278	0.998
South East	1108	0.182	0.188	0.972
South West	700	0.213	0.200	1.064
Wales	410	0.222	0.219	1.015

Source: Office for National Statistics and Family Resources Survey, Department for Work and Pensions

Table 8 Benchmarking results for BHC model, 2011/2012

Country/GOR	Number of MSOAs	Aggregated survey total	Model total	Ratio of survey to model total
North East	340	0.197	0.203	0.973
North West	924	0.179	0.179	1.004
Yorkshire	692	0.190	0.184	1.032
East Midlands	573	0.182	0.184	0.988
West Midlands	735	0.190	0.192	0.991
East	736	0.159	0.160	0.993
London	983	0.160	0.153	1.050
South East	1108	0.149	0.152	0.981
South West	700	0.178	0.163	1.094
Wales	410	0.209	0.205	1.022

Source: Office for National Statistics and Family Resources Survey, Department for Work and Pensions

8. Discussion

Appendix 1 of the 2007/08 technical report summarised detailed development of poverty models for both BHC and AHC income variables for three time periods – 2004/05, 2006/07 and 2007/08. Following this work, it was felt that the best potential for publication for the 2007/08 period lay in the AHC model as this had shown stability over all three of these time periods whereas for BHC this was less clear. For 2011/12, it is felt that now the BHC has attained sufficient stability and thus both the AHC and BHC models have been included in the publication. The models chosen here are among the best for 2011/12. The AHC model explains 82.3 per cent of total between area variability (compared with the intercept only null model), and the BHC 94.2 per cent. For the AHC model, this is a decrease from the model presented for 2007/08, though the change is not large enough to cause concern. In addition, the model diagnostics all (apart from the AHC area level residual plot) perform satisfactorily for both AHC and BHC considering that the direct estimates are obtained from very small sample sizes and that half of the MSOAs have no sample at all.

Validity of the model output estimates has been tested in two ways. A plot of the values at MSOA level for the income domain measure of the Index of Multiple Deprivation 2015 (restricted to England) against the model estimates of poverty shows a correlation coefficient of -0.84 for AHC and -0.85 for BHC. This IMD measure effectively represents the number of the MSOA population in families in receipt of social benefits. A further plot is given of the poverty estimates against the already published MSOA model-based estimates of AHC and BHC household mean income. The observed plots accords with the intuitive expectation of a negative relationship, giving a correlation coefficient of -0.63 for AHC and -0.70 for BHC. When separated out into London and non-London MSOAs, the correlations are higher again. Both validity tests have added confidence to the validity of the AHC estimates.

This report shows that both the AHC and BHC models are well specified, performing well in terms of explanatory power, estimate precision and distinguishability between areas. It also performs well in other time periods. Therefore, the AHC and BHC estimates from the models presented for 2011/12 are published as Experimental Statistics.

9. References

Brown, G., Chambers, R., Heady, P., Heasman, D. (2001).
Evaluation of Small Area Estimation Methods – An Application to Unemployment Estimates from the UK LFS. Proceedings of Statistics Canada Symposium in 2001.

Chambers, R. and Tzavidis, N. (2006).
M-quantile models for small area estimation. *Biometrika*, 93, 255-268.

Elbers, C., Lanjouw, J. O. & Lanjouw, P. (2003).
Micro-level estimation of poverty and inequality. *Econometrica*, 71, 355-364.

Heady, P., Clarke, P., Brown, G., Ellis, K., Heasman, D., Hennell, S., Longhurst, J., Mitchell, B. (2003).
Small Area Estimation Project Report. Model-Based Small Area Estimation Series No.2, ONS Publication.

Longhurst, J., Cruddas, M., Goldring, S., Mitchell, B. (2004).
Model-based Estimates of Income for Wards, 1998/99: Technical Report. Published in Model-Based Small Area Estimation Series, ONS Publication.

Longhurst, J., Cruddas, M., Goldring, S. (2005).
Model-based Estimates of Income for Wards, 2001/02: Technical Report. Published in Model-Based Small Area Estimation Series, ONS Publication.

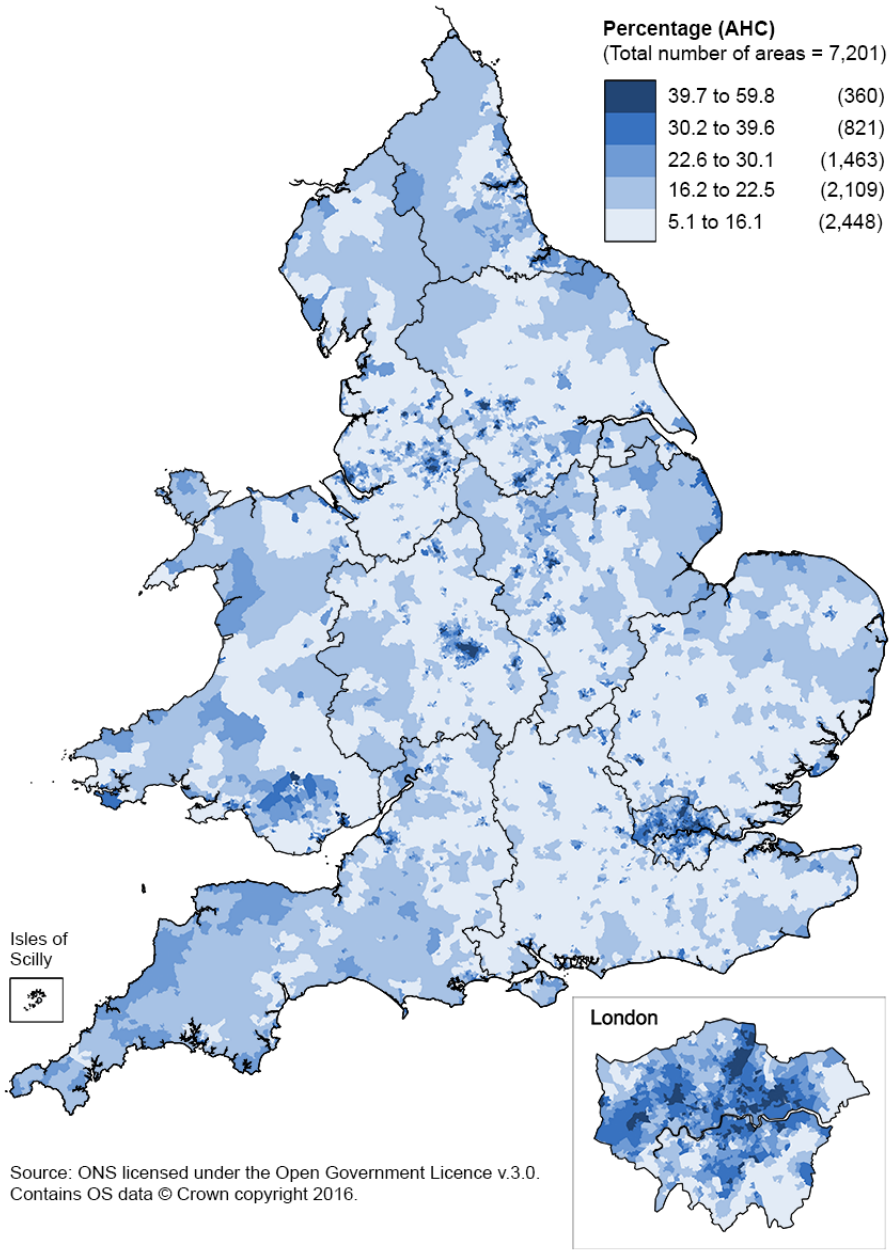
Molina, I, and Rao, J.N.K. (2010).
Small area estimation of poverty indicators. *Canadian Journal of Statistics, 38, 369-385.*

Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2008).
M-quantile models with application to poverty mapping. *Statistical Methods and Applications*, 17, 393-411.

Tzavidis, N., Marchetti, S., and Chambers, R. (2010).
Robust estimation of small area means and quantiles. Australian and New Zealand Journal of Statistics, 52, 167-186.

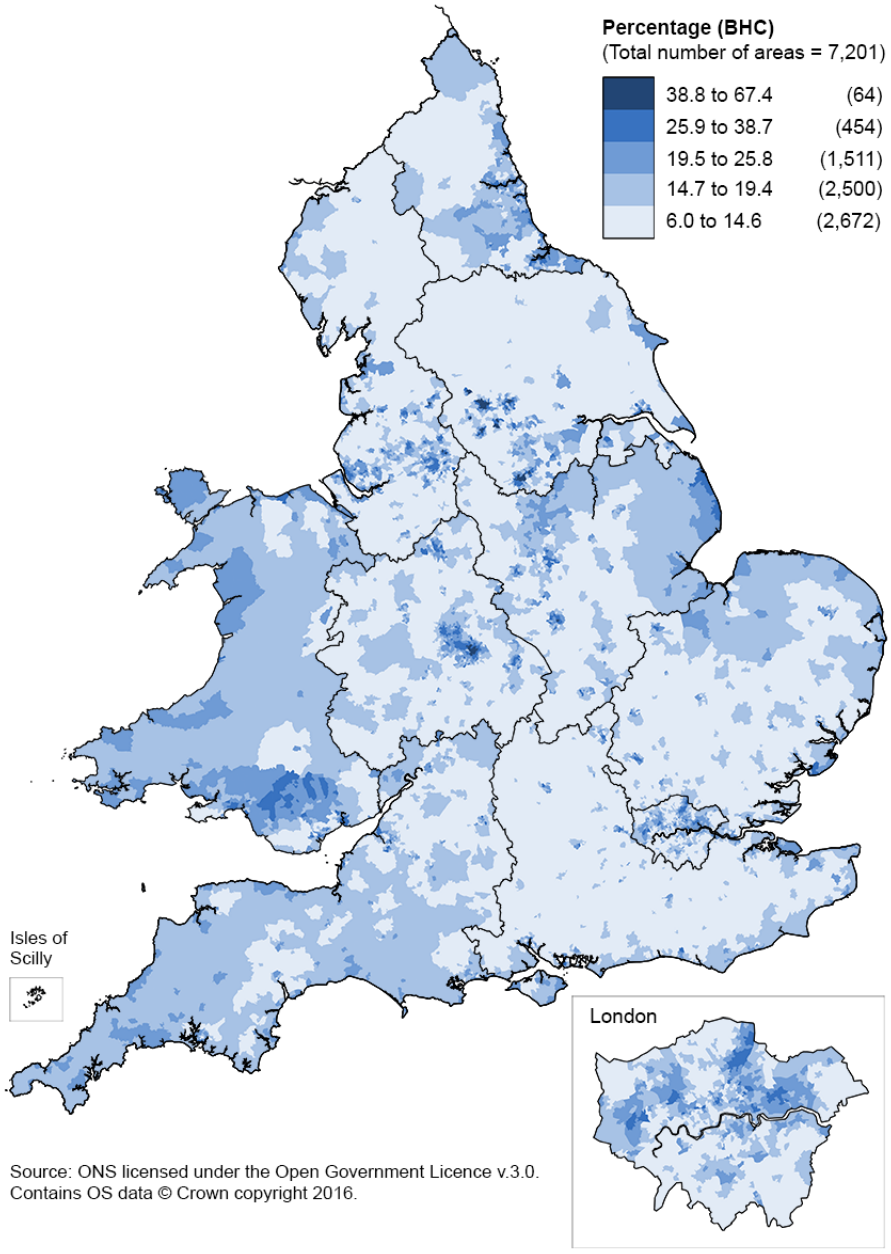
Appendix A: Maps

Map 1 **Households with net equivalised (OECD) income (AHC) < 60% median, 2011/2012 MSOA estimates (calibrated)**



Map 2

Households with net equivalised (OECD) income (BHC) < 60% median, 2011/2012 MSOA estimates (calibrated)



Appendix B: Auxiliary data sources and covariates

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

- Census, 2011
- Department for Work and Pensions benefit claimant counts, August 2011
- Valuation Office Agency Council Tax Bandings, 27 March 2011
- Her Majesty's Revenue and Customs, Child Tax Credit and Working Tax Credit, 2011
- Communities and Local Government, Change of ownership by dwelling price, 2009
- Regional/country identification variable.

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

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

The model selection process for the 2011/12 small area income estimates used variables that were relevant to the time period, so some of the DWP and HMRC variables in Tables B2 and B5 are calculated from the benefits data that were available in 2011/12. The following benefits from these tables have since been replaced with other benefits:

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

B.1 Census Data 2011

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

Table B1: Variables considered for inclusion in modelling income, Census 2011

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

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

B.2 DWP Benefit Data 2011

The DWP benefit data obtained were in the format of counts for each benefit type by MSOA. These counts were transformed into proportions using MSOA population estimates, mid-2011.

Table B2 lists the different DWP variables considered for inclusion in the models as well as the population estimate used as a denominator.

Table B2: Variables considered for inclusion in modelling income, DWP benefit claimant counts 2011

Variable name	Label	Denominator
ISPTOTAL	Proportion of people aged 16 and over claiming Income Support	age16ov
IBSDPTOTAL	Proportion of people aged 16 and over claiming Incapacity Benefit/Severe Disablement Allowance	age16ov
JSAPTOTAL	Proportion of people males aged 16 to 64 and females aged 16 to 59 claiming Job Seekers Allowance	age16-59/64
PCPTOTAL	Proportion of people aged 60 and over claiming Pension Credit	age60ov
PCGEO	Proportion of people aged 60 and over claiming Pension Credit: Guarantee Element Only	age60ov
PCSEO	Proportion of people aged 65 and over claiming Pension Credit: Saving Element Only	age65ov
PCGESE	Proportion of people aged 65 and over claiming Pension Credit: Guarantee and Saving Element	age65ov
DLATOTAL	Proportion of people claiming Disability Living Allowance	ageall
DLAMAL	Proportion of people claiming Disability Living Allowance: Mobility Award Lower	ageall
DLAMAH	Proportion of people claiming Disability Living Allowance: Mobility Award Higher	ageall
DLACAL	Proportion of people claiming Disability Living Allowance: Care Award Lower	ageall
DLACAM	Proportion of people claiming Disability Living Allowance: Care Award Middle	ageall
DLACAH	Proportion of people claiming Disability Living Allowance: Care Award Higher	ageall

B.3 Regional and Country identification variable

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

Table B3: Regional variables included in modelling income

Variable name	Country/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

B.4 HMRC Child Tax Credit and Working Tax Credit Data 2011/12

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

Table B4 lists the HMRC variables considered for inclusion in the models.

Table B4: Variables considered for inclusion in modelling income, HMRC Child Tax Credit and Working Tax Credit Data 2011/12

Variable name	Label
TFAMTC	Families Receiving; Tax Credit
FAMWKTC	Families in work receiving; Tax Credit
LPWKTC	Lone-parent families in work receiving; Tax Credit
FAMWKCTWT	Families in work receiving; Child Tax Credit and Working Tax Credit
FAMWKAFE	Families in work receiving; Child Tax Credit above the family element
FAMWKBFE	Families in work receiving; Child Tax Credit family element and below
FAMWKWT	Families in work receiving; Working Tax Credit only
FAMOUTCT	Families out of work receiving; Child Tax Credit
LPOUTCT	Lone-parent families out of work receiving; Child Tax Credit
CPOUTCT	Couple families out of work receiving; Child Tax Credit

B.5 Valuation Office Agency council tax band data

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

The Council Tax bands for England and Wales are not consistent; therefore separate covariates are defined for England and Wales. In Wales, some MSOAs have very high concentrations at one end of the range of tax bands, causing model instability. The final covariates considered for inclusion in the model are shown in Table B5.

Table B5: Variables considered for inclusion in modelling income, VOA Council Tax Bands, 2011

Variable name	Label
Engabc	Proportion of dwellings in English Council Tax bands A, B and C
Engdef	Proportion of dwellings in English Council Tax bands D, E and F
Engghi	Proportion of dwellings in English Council Tax bands G, H
Walabc	Proportion of dwellings in Welsh Council Tax bands A, B and C
Waldef	Proportion of dwellings in Welsh Council Tax bands D, E and F
Walghi	Proportion of dwellings in Welsh Council Tax bands G, H and I

B.6 Department for Communities and Local Government Change of ownership by dwelling price, 2009

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

Table B6 lists the CLG variables considered for inclusion in the models.

Table B6: Variables considered for inclusion in modelling income, CLG change of ownership data 2009

Variable name	Labels
TRNS	Transactions by Dwelling Type; Total Sales
PLQ	Price Indicators for All Dwellings; Lower Quartile
PMED	Price Indicators for All Dwellings; Median
PUQ	Price Indicators for All Dwellings; Upper Quartile
PMEAN	Price Indicators for All Dwellings; Mean
OUTC	Number of Outliers; more than £20m

Appendix C: Data Preparation

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