Dadansoddi ar gyfer Polisi

YMCHWIL GYMDEITHASOL Y LLYWODRAETH

GSR

GOVERNMENT SOCIAL RESEARCH

Analysis for Policy

Llywodraeth Cymru
Welsh Government

www.cymru.gov.uk

Ymchwil gymdeithasol

Social research

Number: 46/2015

Understanding Wales at the neighbourhood level: maximising the performance of small area estimation

Understanding Wales at the small area level: maximising the performance of small area estimation

Dr Adam Whitworth and Eleanor Carter

University of Sheffield

Views expressed in this report are those of the researchers and not necessarily those of the Welsh Government

For further information please contact:

Chris McGowan

National Survey team

Welsh Government

Cathays Park

Cardiff CF10 3NQ

029 2080 1067

Email: chris.mcgowan@wales.gsi.gov.uk

Welsh Government Social Research, 26 August 2015

ISBN 978-1-4734-4488-1

Table of contents

List of Figures

List of Tables

Executive summary

1. Small area estimation (SAE) describes a range of alternative statistical techniques for the estimation of survey data down to small area level where those data are not otherwise available at these smaller spatial scales – income, fear of crime, healthy lifestyles, digital engagement to name but a few.

2. This project follows on from a recent Welsh Government project around the potential role of SAE to create new small area data from existing survey datasets in order to offer otherwise unobtainable spatial knowledge for policy analysis and decision-making (Whitworth et al., 2015). This current project focuses particularly on methodological experimentation and learning in order to best inform the Welsh Government around the future potential of SAE in its work and the most fruitful ways to approach any such future SAE work. As with the previous work, this project focuses on the iterative proportional fitting (IPF) spatial microsimulation approach to small area estimation.

3. The project incorporates several innovative elements in order to maximise methodological learning:
    - it compares the performance of alternative approaches to the tailored selection of survey cases based on differing area characteristics (deprivation, geodemographic type, urbanity/rurality, region) and further explores the effect of combinations of these area characteristics;
    - it examines the impact of post-estimation constraining techniques to ensure that estimated counts sum to known values at higher spatial scales;
    - it includes outcome variables known at small area level from the Census 2011 (poor health and Welsh speaking) in order to allow unusually robust external validation of the estimates at the target small area spatial scales; and
    - it compares the impact on resulting small area estimates when base surveys of smaller size are relied upon.

4. Overall the project demonstrates that where outcomes are able to be modelled with a reasonable degree of predictive power the IPF approach is a viable methodological approach for the production of accurate small area estimates and accompanying credible intervals. Careful thought needs to be placed on the specification of the small area estimation approach in order to maximise the quality of the resulting small area estimates, including consideration of the optimal combination of individual and area-level constraints. These will vary depending upon the source of the variation in the outcomes being estimated and will affect both the estimation of the point estimates and credible intervals. Post-estimation constraining to known ‘true’ values at higher scales is shown by this project to be effective, and this may be considered as an additional step to further enhance resulting small area estimates.

5. The project is divided into five work strands:
    - Work strand one: to produce small area estimates of the percentage of adults who use the internet, have access to the internet at home, speak Welsh, and are in poor health using the National Survey for Wales 2013-14 results;
    - Work strand two: to explore the potential for, and impact of, tailored survey case selection prior to the IPF estimation on the estimation of these four outcomes;
    - Work strand three: test the accuracy of the small area estimates of Welsh speaking and poor health by validating externally against values known at those small scales from Census 2011;
    - Work strand four: explore the potential for, and impact of, post-estimation constraining of those small area estimates to known 'true' values at higher geographical levels;
    - Work strand five: to explore the potential to produce estimates of the percentage of adults who smoke, drink above guidelines, are obese, and are physically active, using the Welsh Health Survey data from 2008 to 2013.

6. In addition to producing new estimates of internet access and internet use at both Lower Layer and Middle Layer Super Output Area scales, **the first two work strands** offer important methodological learning through the evaluation of alternative IPF specifications based respectively on individual factors and a combination of individual and area factors. It was possible to construct predictive models of acceptable explanatory power for all four target outcome variables: internet access, internet use, Welsh speaking and poor health. In general, a combination of individual and area factors is most effective, although the benefits of area factors vary dramatically dependent upon the source of variation seen in the outcome variable. Location, for example, is key to explaining Welsh speaking but is relatively unimportant in explaining poor health.

7. **Work strand three** focuses on the key task of validation, particularly in terms of the external validation of the small area estimates produced in work strands one and two against known 'true' external data. The internet outcomes are by necessity validated externally at the higher local authority scale and validate well at this aggregated scale. The inclusion of Welsh speaking and poor health as target outcome variables enables the project to provide valuable insights into the potential quality of small area estimates by enabling robust external validation at the target small area scales using known values from Census 2011. The IPF estimates validate well against those known Census values at the small area level with Pearson's correlation coefficients of 0.92 (percentage of MSOA adults in poor health) and 0.93 (percentage of MSOA adults speaking Welsh) respectively between the 'true' Census values and the IPF estimates. In addition, the IPF estimates out-perform equivalent small area estimates produced in a separate sub-project using an area-level regression approach based on external validation statistics that compare both sets of resulting small area estimates with

known ‘true’ values from the Census 2011. The main difference between the two approaches is that the IPF approach is able to incorporate both individual and area-level factors whilst this regression modelling approach is able to include only factors relating to the areas in which survey respondents live and not relating to the characteristics of the survey individuals themselves.

8. Once small area estimates have been produced, **work strand four** explores the potential of alternative approaches to constraining these estimates to known ‘true’ values at higher geographical scales such that their estimated totals come to match. Constraining approaches are only occasionally used in the small area estimation literature, in part due to the assumptions that come with them, but two of the three constraining approaches tested within work strand four do show overall benefits in terms of the quality of the final small area estimates and suggest that post-estimation constraining may be a useful complement to small area estimation work.

9. Whilst the analyses in strands one to four are based on the National Survey for Wales 2013-14, **work strand five** offers a stand-alone exploration of the viability of producing healthy lifestyle estimates at the somewhat larger Upper Layer Super Output Area (USOA) scale based on the Welsh Health Survey results from 2008 to 2013. IPF estimates are created in relation to the percentage of Welsh adults in each USOA affected by four healthy lifestyle outcomes – smoking, obesity, adequate physical activity, and alcohol consumption above guidelines – and are compared to existing USOA estimates generated directly from the Welsh Health Survey. Analyses also explore the impact of basing the small area estimation on smaller base survey datasets. None of these four healthy lifestyle outcomes is able to be well predicted by underlying regression models and this places the small area estimation of these outcomes on vulnerable foundations as it reduces their likelihood of being able to deliver acceptably accurate small area estimates. Nevertheless, work strand five shows that acceptable small area estimates can potentially be produced from relatively weak predictive foundations such as these and this is seen in the small area estimates for smoking and obesity in particular that validate well against the direct survey estimates. In terms of possible future work, one could attempt the future estimation of weakly predicted outcomes if it were possible to perform robust external validation the findings and if it were accepted that there is a larger than usual risk of the results validating poorly when working with weakly predicted outcomes. Further research could explore the conditions under which the conversion between underlying model power and the quality of the external validation is maximised in order that clearer guides can be established around the likely effectiveness of small area estimation from underlying models of varying predictive power.

Introduction

Why small area estimation?

Small area estimation (SAE) refers to a set of methodological techniques to estimate survey data down to small area levels where those data are not otherwise attainable. Examples in the UK include variables related to healthy lifestyles, fear of crime, attitudes, well-being or income. All of these variables are of policy interest but they are not available in the UK at small geographical scales from other sources such as the Census 2011 or government administrative data. Collecting new survey data of sufficient small area sample sizes to give acceptable estimates at small geographical scales for the whole of Wales would be prohibitively expensive. SAE techniques have the potential to provide spatially detailed information to guide policy decisions, maximising the value of existing survey data investments and minimising the need for expensive large-scale survey data collection.

Project aims

A recent project for the Welsh Government used the iterative proportional fitting (IPF) spatial microsimulation approach to SAE to begin to explore the potential that SAE may offer policy makers in both central and local government across Wales (Whitworth et al., 2015). This previous study applied IPF to the National Survey for Wales 2012-13 data combined with small area covariate data from the Census 2011 in order to produce MSOA level estimates of the percentage of adults aged 16+ who: use the internet; are experiencing financial difficulties; feel unsafe in the local area after dark; are satisfied with their GP care; are highly satisfied with their local area; and are highly satisfied with the performance of the Welsh Government. All MSOA estimates were provided with accompanying credible intervals to provide a sense of the uncertainty around the point estimates and the validation work shows that the estimates in general validate well. Further exploratory analyses within this previous project also highlighted the potential to produce estimates down to the smaller LSOA level scale as well as to produce local authority estimates from smaller base surveys.

The current project furthers this previous work across five linked work strands that collectively aim not only to produce new small area estimates but also to explore the impact of differing methodological innovations on the refinement of the small area estimation. The five work strands of the current project are:

- **Work strand one**: to produce small area estimates (LSOA and MSOA level) of the percentage of adults who report that they use the internet, have access to the internet at home, speak Welsh, and are in poor health using the National Survey for Wales 2013-14;
- **Work strand two**: to explore the potential for, and impact of, tailored survey case selection on the small area estimation of these four outcomes;
- **Work strand three**: to test the accuracy of the small area estimates by validating externally against known ‘true’ values, including at the small area level for the Welsh speaking and poor health outcomes given their collection within Census 2011;

- **Work strand four**: to explore the potential for, and impact of, post-estimation constraining techniques in order to adjust resulting small area estimates to known 'true' values at higher geographical levels;
- **Work strand five**: to explore the potential to produce small area estimates (USOA level) of the percentage of adults who smoke, are obese, are physically active and drink above guidelines, using the Welsh Health Survey results from 2008 to 2013.

As outlined above, each work strand is designed with a specific purpose in mind such that taken together they both produce new small area estimate as well as offering empirically grounded insights and recommendations to support Welsh Government in any desired future small area estimation work. Table 1 offers a summary outline of the aims, outcome variables, survey data sources and spatial scales across each of the five work strands of the project.

Table 1: Summary of the data sources, outcome variables and geographical scales across the five work strands

Survey data source	Outcome variable to be estimated	Strand 1: Single-level IPF at LSOA & MSOA scales	Strand 2: Multi-level IPF at MSOA scale	Strand 3: External validation of LSOA and MSOA estimates	Strand 4: Post-estimation constraining of MSOA estimates	Strand 5: Healthy lifestyle estimates at USOA scale
National Survey for Wales 2013-14	% of Welsh adults using the internet[1]	X		X	X	
	% of Welsh adults with access to the internet[2]	X		X		
	% of Welsh adults who can speak Welsh[3]	X	X	X		
	% of Welsh adults in poor health[4]	X	X	X	X	
Welsh Health Survey 2008 to 2013	% of Welsh adults who smoke daily or occasionally					X
	% of Welsh adults who drink alcohol beyond guideline levels[5]					X
	% of Welsh adults who are obese according to their BMI					X
	% of Welsh adults who are physically active[6]					X

[1] 1= 'Personal use of the internet at work, home or elsewhere'; 0= 'No personal use of the internet at work, home or elsewhere'
[2] 1= 'Household has access to the internet'; 0= 'Household does not have access to the internet'
[3] 1= 'Can speak at least a little Welsh'; 0= 'No cannot speak Welsh, or can only say a few words'
[4] 1= 'Bad general health' or 'Very bad general health'; 0 = 'Very good general health' or 'Good general health' or 'Fair general health'
[5] 1= 'Maximum daily alcohol consumption exceeded guidelines at least once in the past week'; 0= 'Never exceeded alcohol consumption guidelines in the past week'
[6] 1= 'Do at least 30 minutes of moderate/vigorous exercise on 5 or more days per week'; 0='Do not do at least 30 minutes of moderate/vigorous exercise on 5 or more days per week'

Methodology

Data

SAE is a set of diverse techniques to estimate survey variables of interest down to small area scales where those outcomes are not otherwise available at those smaller geographies. In order to do so the survey must offer the outcome variable of interest, and both the survey data and the small area data (typically but not necessarily sourced from the Census) must offer the same set of explanatory/covariate data. As summarised in Table 1, strands one to four make use of the National Survey 2013-14 results, provided by the Welsh Government with Lower Layer Super Output Area (LSOA) geocodes attached. The estimation of the healthy lifestyle outcomes in strand five makes use of a pooled dataset of annual Wealth Health Survey results from 2008 to 2013 inclusive, provided by the Welsh Government with LSOA geocodes again attached. The small area covariate data for all strands come from the UK Census 2011 and are sourced from the NOMIS data portal website[7]. After cleaning and recoding the National Survey provides a base survey of 13,566 cases for the IPF across strands one to four, whilst the pooled Welsh Health Survey data provide a base survey of 62,269 survey cases.

Spatial scales

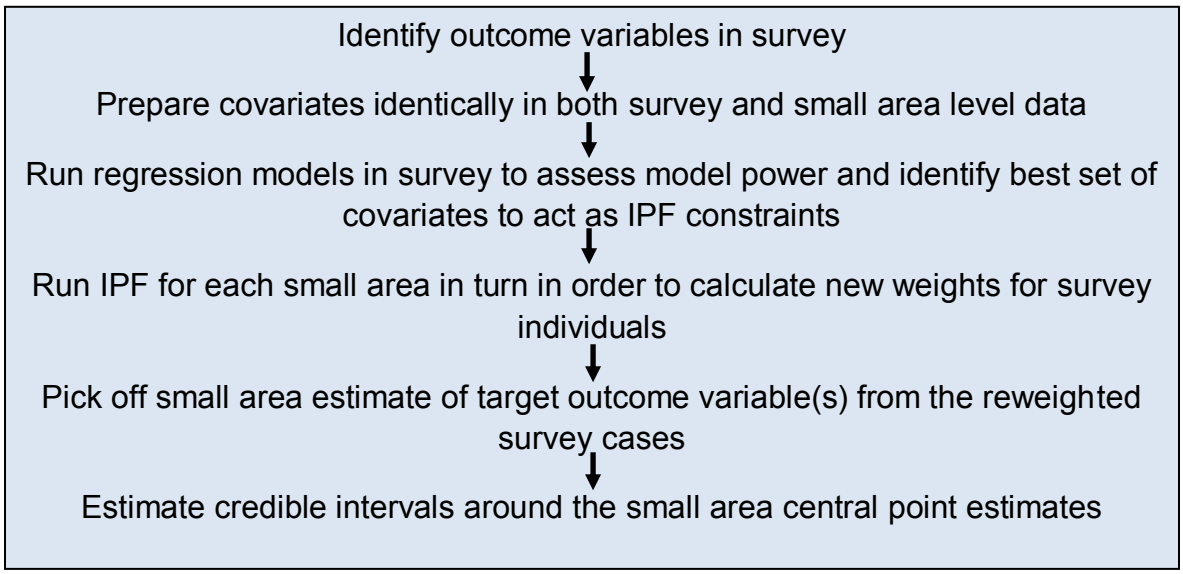
As summarised in Table 1, the project produces and evaluates small area estimates at a range of spatial scales below those of local authority areas. Strand one creates estimates at both Middle Layer Super Output Area (MSOA) and Lower Layer Super Output Area (LSOA) whilst strands two and four focus in different ways on refining the MSOA estimates (in work strand two via the pre-estimation tailoring of case selection and in work strand four via post-estimation constraining). The average population size of MSOAs in England and Wales is 7,860 whilst for LSOAs it is 1,630 (ONS, 2012; 2013). For the healthy lifestyles estimates in strand five the estimates are produced at the somewhat larger Upper Layer Super Output Area (USOA) geography as this is the scale at which comparable direct survey estimates for these outcomes have been created. USOAs have an average population size of 26,700.

Understanding iterative proportional fitting

Small area estimation represents not one but many alternative methodological approaches to estimate survey data down to small geographical scales where those variables are not otherwise obtainable at those fine spatial scales[8]. This project makes use of a widely used spatial microsimulation approach known as iterative proportional fitting (IPF). Figure 1 below summarises the main steps in the IPF method.

[7] http://www.nomisweb.co.uk/census/2011
[8] For helpful methodological summaries see Whitworth (2013), Marshall (2010) and Rahman (2008).

Figure 1: Overview of key steps in the IPF method



Preparing the data

The initial tasks in an IPF approach are to identify the potential outcome variable(s) of interest in the survey and the set of explanatory factors in the survey to best explain those outcomes. Outcome variables in IPF are typically either continuous or binary variables. In this project, all outcome variables are binary in nature.

As detailed below, IPF operates by reweighting survey cases so that in the aggregate the survey cases match the characteristics of the target small area across the set of selected explanatory/constraint variables. Regression models are important to identify the optimal set of explanatory/constraint factors to predict the target outcomes and to assess the combined predictive power of those factors in accounting for the variation in the outcomes. In identifying these explanatory variables in the survey an important consideration is the ability to also source them at the small area level. This represents an inevitable restriction on IPF, as with all other SAE approaches.

The power of these underlying regression models (assessed typically via the R-squared or, in logit models, pseudo R-squared values) provides information about the strength of the modelled relationships seen and gives an initial guide as to the likely accuracy of the final small area estimates. There is no strict one-to-one linkage between the predictive power of these models and the final validity[9] of the small area estimates however. Nevertheless, one does desire reasonably powerful models in order to increase confidence in being able to produce valid estimates of the outcomes at the small area level: if models are weak then survey characteristics associate less strongly with the target outcomes. This means that even effectively reweighted cases are less likely to deliver valid small area estimates and more likely to be volatile across different survey samples given that the estimation is based on relationships between constraints and outcome(s) that are less systematic. No strict rules exist as to desirable minimum thresholds in terms of model power and in

[9] 'Validity' here refers to their close correlation to 'true', but usually unknown, values at the small area level.

practice this is a gradual sliding scale of more- or less-desirable values: as a rule of thumb an R-squared value of greater than around 20%-30% at a minimum seems advisable.

Once the most powerful and parsimonious models have been identified for each outcome variable in the survey the selected explanatory/constraint variables can be prepared. Explanatory variables are prepared as binary variables in the survey dataset and identical definitions of these same explanatory variables are prepared as small area totals in the small area covariate dataset. For example, a variable called 'male' would be coded 1 for males and 0 for females in the survey whilst a variable called 'female' would also be created and would be coded 1 for females and 0 for males. In terms of the small area covariate data IPF requires aggregate totals at the small area level (e.g. total number of males and females in each small area) and it is necessary that the different small area totals sum to the same population totals. In this project the Census 2011 data for Welsh small areas contained some variation across the different Census tables in the total number of Welsh adults that they summed to. This is most apparent where tables relate to different universes of sampled individuals, particularly whether students were or were not included in a particular Census table. As a final step it was necessary therefore to ensure that the small area covariate totals of the constraint variables were adjusted so that that they summed to the same value: the simple age-sex band Census table was taken as the 'true' small area population value for this purpose and all other small area totals were adjusted to meet this.

Running the IPF

Once the survey and small area covariate datasets have been prepared, the IPF can be implemented. The IPF takes each small area in turn and fits the survey cases as effectively as possible to the multi-dimensional profile of the target small areas across each of the specified constraint variables. The IPF achieves this by sequentially reweighting the survey cases across each constraint in turn, based on the extent to which the (re)weighted sum of each constraint in the survey file matches the small area total for that constraint.

Ballas and Anderson (in Whitworth, 2013) provide an example of the IPF methodology applied, and an adjusted version of this example is shown below. An example segment of annotated Stata syntax for the IPF is included in the Technical Appendix. Intuitively, one can understand in this example that the aggregated household survey data shown in Table 3 has 'too many' people in it for the target small area shown in Table 2: the small area being reweighted to contains 7,840 earners according to Table 2 but the survey file contains 12,310 earners. Overall, therefore, the survey cases need to have their weights reduced in size in order that the weighted total of earners in the survey file comes to match the number of earners in this target small area. More precisely, however, the IPF seeks to match the number of earners across each separate grouping of this constraint – zero earners, one earner, two earners, three plus earners – and the ratios between the weighted survey totals in Table 3 and the small area totals in Table 2 are not equal across these categories. As a consequence, the IPF will reduce the weight of each survey

case differently according to the degree of mismatch as captured in the ratio between the two totals. If a survey case contains zero earners, for example, then this household will have its weight multiplied by 0.73 (i.e. 3,970/5,440 = 0.73) whilst if a survey household contains two earners then it will be down-weighted to a greater extent by having its weight multiplied by 0.46 (1,420/3,090=0.46).

Table 2: Small Area figures for number of earners (i.e. individuals in paid employment) derived from Census 2011 for the first MSOA in Wales

MSOA	Number of individuals	Number of earners = 0	Number of earners = 1	Number of earners = 2	Number of earners = 3+
MSOA1	7840	3970	2210	1420	240

Table 3: Small Area figures for number of earners (i.e. individuals in paid employment) derived from weighted sum of National Survey dummy variables

MSOA	Number of individuals	Number of earners = 0	Number of earners = 1	Number of earners = 2	Number of earners = 3+
MSOA1	12310	5440	3260	3090	520

To begin the IPF process all individuals are given their adult weight as provided within the survey. For each constraint in turn the weights for each individual in the survey are then adjusted using the formula below:

*New weight = Previous Weight ** (*MSOA constraint total / Weighted survey constraint total*)

Table 4 shows a worked example of the calculations for this first constraint – the number of earners. For each survey case the original survey weight is adjusted so that the survey sample fits the Census data on this one dimension and a new weight is calculated. In this example the effect of the reweighting on this constraint is to reduce the size of the weights and, more specifically, to reduce them to differing degrees for each earner category dependent on their differing ratios with the Census small area total. This new weight then becomes the starting weight for the fitting on the next constraint and this reweighting process continues across all of the constraints in turn.

In order to ensure that the weights are refined adequately until they stabilise, once the IPF has passed over the constraints once it then loops back to the start and moves sequentially over the same set of constraints again. There is no agreement within the literature about how many times the IPF should be set to iterate around the full set of constraints, with Ballas et al (2005) recommending 5-10 times and Anderson (2007) suggesting 20 times. In this project we found that 10 iterations were enough to produce stable weights and successful fitting. Once final IPF weights have

been created for each survey case for that small area then the process moves on to the next small area and repeats until all small areas have been processed.

Table 4: First four survey individuals with adjusted weights after fitting to constraint 1

Survey Case	Number of earners	Initial survey weight	New weight after fitting to constraint 1 (number of earners)
1	1	51.2	= 51.2 * (2210/3260) = 34.7
2	0	76.3	= 76.3 * (3970/5440) = 55.7
3	2	33.7	= 33.7 * (1420/3090) = 15.5
4	1	125.3	= 125.3 * (2210/3260) = 84.9
..	..	…	..

For each small area the end result is a reweighted version of the survey file where the reweighted survey individuals represent the 'fractional existence' of that kind of individual in that small area and where the reweighted survey file as a whole can be understood as a synthetic population micro-dataset for the small area. The estimated small area values can then be picked off as weighted values of the target outcome variable from the survey cases for each small area using the final IPF weights of each survey case for each target small area. As all target outcome variables are binary variables in this project, weighted sums are calculated to give an estimate of the total number of adults in the small area affected (e.g. the total number of adults estimated to use the internet in each Welsh MSOA). Given known adult population totals these sums can easily be converted to estimated percentages.

Calculating the credible intervals

A final task in the process is to provide a sense of likely uncertainty around the central small area point estimates. Spatial microsimulation methods such as IPF typically do not provide such intervals alongside their point estimates. For this project, however, it was considered important to provide a sense of uncertainty around the point estimates this is done through the development of an approach used to calculate credible intervals within the statistical literature (Heady et al., 2003; Bajekal et al., 2004; Pickering et al., 2004).

As discussed above, the identification of the optimal set of constraints for the IPF is based on single-level regression models and such models produce clear and widely understood summary measures of model power (R-squared or pseudo R-squared). However, given that the data (and the IPF process itself) show a multilevel structure of individuals (level one) nested inside target small areas (level two) it is possible to understand these as hierarchical data structures suitable for multilevel regression models. Unlike single-level models, in a multilevel regression structure the error variance is partitioned across the different levels in the model. In the empty multilevel model (i.e. the model with no explanatory variables) this is represented by the

intraclass correlation coefficient (ICC) which is a measure of the error variance accounted for at the area level compared with at the individual level. In the full multilevel model (i.e. the model that includes explanatory variables) the residual error variance at the different levels is reported after having accounted for the variables in the model. A comparison of those error variances between the empty and the full multilevel models indicates the extent to which the explanatory factors are able to account for the variance in the outcome at each level of the model.

Given that the emphasis in the small area estimation process is on the estimates at area level, it is the residual variance within the level two (i.e. area-level) error term that is of central interest in terms of the credible intervals as this is the key indicator of remaining between-area uncertainty. In order to compute the credible intervals around the central point estimates we therefore follow the following steps:

- take the central IPF point estimate and express it in terms of log odds;
- calculate the distribution of the unexplained between-area error term that is to be incorporated in order to give a sense of uncertainty around the central IPF point estimate: this is calculated with a mean error of zero and a variance as given by the estimate (also expressed in terms of log odds) of the residual variance on the level two (i.e. between-area) error term within the full multilevel model for the outcome at the target spatial scale (i.e. MSOA or LSOA);
- take 10,000 separate randomly drawn values from this level two error distribution[10], add each one to the central IPF point estimate and convert all 10,000 resulting values from log odds to probabilities;
- the values of the 250th and 9,750th largest cases (i.e. the 2.5th and 97.5th percentiles) of this resulting distribution of 10,000 values are taken as the lower and upper limits of the credible intervals around the central IPF point estimate.

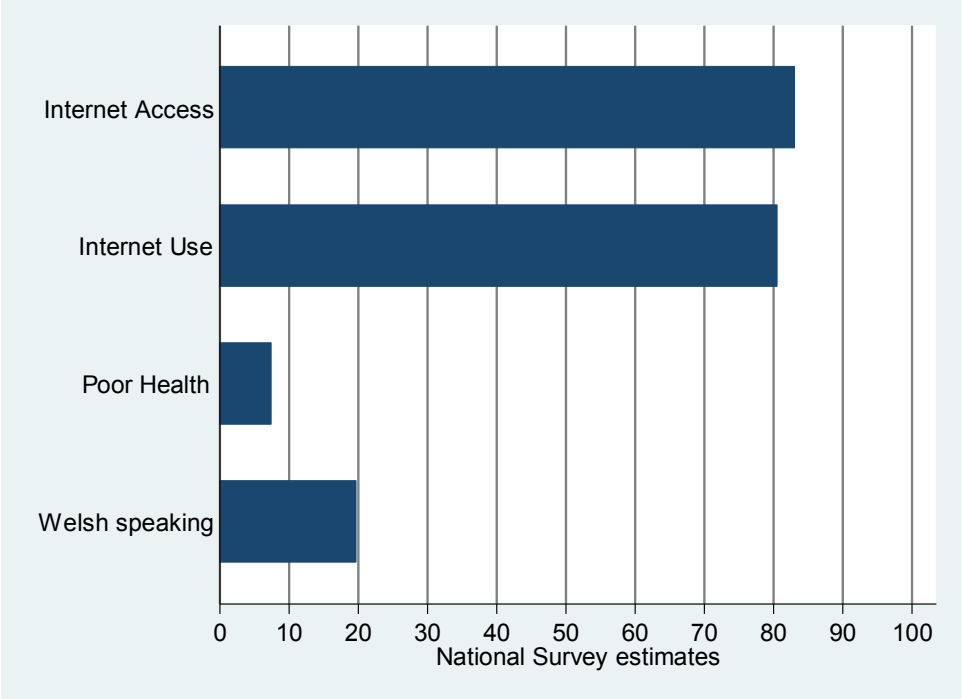
[10] The rnormal() function in Stata is used to create the credible intervals and this is in fact based on standard deviation rather than variance.

Work strand one: Single-level IPF at LSOA and MSOA scales

The first strand of the project produces estimates of four target outcome variables using single-level IPF that incorporates constraints relating to individuals only, and no constraints relating to the areas in which those individuals live. The four outcome variables estimated are self-reported Welsh speaking, self-reported poor health, household internet access and internet use, and these are estimated at both LSOA and MSOA scales. The Welsh speaking and poor health outcomes are selected specifically to enable rigorous external validation of the resulting IPF estimates at the small area scale in work strand three.

Figure 2 below shows the context in which the small area estimation takes place. Across Wales as a whole, weighted direct survey estimates of the National Survey 2013-14 show that around 83 per cent of Welsh adults have access to the internet, a little over 80 per cent use the internet, around 8 per cent self-report that they are in poor health, and a little under 20 per cent self-report that they speak Welsh.

Figure 2: National Survey estimates of the percentage of Welsh adults for each outcome



The first task in the IPF process is to identify the optimal set of constraints for each outcome and to assess their combined predictive capacity. Tables 5 and 6 show the results of the optimal regression models for each of these four outcome variables. These tables therefore show the optimal set of individual-level constraints used in the IPF of each of these four outcome variables. These constraints are incorporated into the IPF in the order shown in these tables, with the strongest predictor acting as the final constraint in the IPF.

Tables 5 and 6 show the log odds, and indication of their statistical significance, for each of those explanatory factors within single-level binary logistic regression models. The final row of each table shows the pseudo R-squared values for each model and this gives an indication of the likely ability to create acceptable small area

estimates. For three of the outcome variables – poor health, internet access and internet use – the pseudo R-squared values are in the order of 35 per cent to 40 per cent and this is considered a solid base from which to perform the small area estimation. For the Welsh speakers outcome, however, the best set of individual-level explanatory factors are able to account for only around 10 per cent of the total variation. This suggests that it may be more difficult to produce acceptable small area estimates of the percentage of adults who speak Welsh when using only individual-level constraints in the IPF in work strand one compared with the three other outcomes.

Table 5: Binary logistic regression models underpinning the single-level IPF of Welsh speaking and poor health in strand one

Welsh Speaking				**Poor Health**		
Explanatory Variable		**Log odds**		**Explanatory Variable**		**Log odds**
National Identity (ref=Welsh only)	Welsh or Welsh British	-0.49**		**Disability (ref=None)**	Has limiting illness	54.22**
	Rest of UK	-1.26**		**Economic Activity (ref=In Work)**	Unemployed	1.54*
	Other	-2.31**			Retired	2.50**
Highest Qualifications (ref=no qualifications)	Level 1	-0.09			Inactive	4.23**
	Level 2	0.37**			Student	0.68
	Level 3	0.54**		**Highest Qualifications (ref=no qualifications)**	Level 1	0.81*
	Level 4+	0.83**			Level 2	0.83*
Country of Birth (ref=England)	NI	0.18			Level 3	0.86
	Scotland	-0.20			Level 4+	0.65**
	Wales	0.40**		**Age-Sex group (ref= Male 16-19)**	Female 16-29	1.82
	Ireland	0.09			Female 30-49	3.40**
	Other	-0.02			Female 50-64	5.05**
Occupational Classification (ref=Higher)	Intermediate	0.02			Female 65+	3.37**
	Routine Manual	-0.38**			Male 30-49	3.35**
	Never Worked	-0.38**			Male 50-64	4.69**
Age-Sex group (ref= Male 16-19)	Female 16-29	0.16			Male 65+	3.26**
	Female 30-49	-0.12		**Tenure (ref=Owner Occupier)**	Social rent	1.62**
	Female 50-64	-0.17			Private rent	1.28*
	Female 65+	0.07			Constant	0.00**
	Male 30-49	-0.35**		**Observations**	13566	
	Male 50-64	-0.36**		**Pseudo-R2**	39.80%	
	Male 65+	-0.12				
Vehicle Access (ref=car access)	No Car	-0.25**				
	Constant	-0.13**				
Observations	13566					
Pseudo-R2	9.57%					
****=p<0.05, *=p<0.1**						

Table 6: Binary logistic regression models underpinning the single-level IPF of internet access and internet use in strand one

Internet Access				Internet Use		
Explanatory Variable		Log odds		Explanatory Variable		Log odds
Household Type (ref=Two adults with kids)	Single Pensioner	-1.90**		Age-Sex group (ref= Male 16-29)	Female 16-29	-0.09
	Married Pensioner	-0.76**			Female 30-49	-0.96**
	Single Working Age	-1.83**			Female 50-64	-1.95**
	Two Adults No Kids	-0.34**			Female 65+	-3.06**
	Lone Parent	-0.88**			Male 30-49	-1.18**
	Other	-0.51**			Male 50-64	-2.22**
Highest Qualifications (ref=no qualifications)	Level 1	0.61**			Male 65+	-3.02**
	Level 2	0.81**		Highest Qualifications (ref=No qualifications)	Level 1	0.91**
	Level 3	1.19**			Level 2	1.04**
	Level 4+	1.40**			Level 3	1.56**
Economic Activity (ref=In Work)	Unemployed	-0.38**			Level 4+	1.70**
	Retired	-0.63**		Economic Activity (ref=In Work)	Unemployed	0.24
	Inactive	-0.35**			Retired	-0.76**
	Student	0.41*			Inactive	-0.50**
Age-Sex group (ref= Male 16-29)	Female 16-29	-0.14			Student	0.73**
	Female 30-49	0.21		Tenure (ref=Owner Occupier)	Social rent	-0.42**
	Female 50-64	-0.31*			Private rent	-0.11
	Female 65+	-1.27**		Household Type (ref=Two adults with kids)	Single Pensioner	-1.00**
	Male 30-49	-0.04			Married Pensioner	-0.41**
	Male 50-64	-0.76**			Single Working Age	-0.81**
	Male 65+	-1.37**			Two Adults No Kids	-0.35**
Vehicle Access (ref=No Car)	Access Car	0.79**			Lone Parent	-0.25
Occupational Classification (ref=Higher)	Intermediate	-0.45**			Other	-0.56**
	Routine Manual	-0.82**		Vehicle Access (ref=No Car)	Access Car	0.82**
	Never Worked	-0.90**		Occupational Classification (ref=Higher)	Intermediate	-0.48**
Tenure (ref=Owner Occupier)	Social rent	-0.63**			Routine Manual	-1.01**
	Private rent	-0.38**			Never Worked	-1.06**
	Constant	2.63**			Constant	3.32**
Observations	13566			Observations	13566	
Pseudo-R2	35.90%			Pseudo-R2	40.60%	

In order to calculate the credible intervals around the central IPF small area point estimates, multilevel binary logistic models with the same set of explanatory factors are implemented. In these multilevel models survey individuals are at level 1 and the target scale for the small area estimates (i.e. either LSOAs or MSOAs) are at level 2. Table 7 presents the standard deviations of these residual level 2 error terms for these multilevel models expressed with log odds as the units. These values are key to the width of the resulting credible intervals around the small area point estimates.

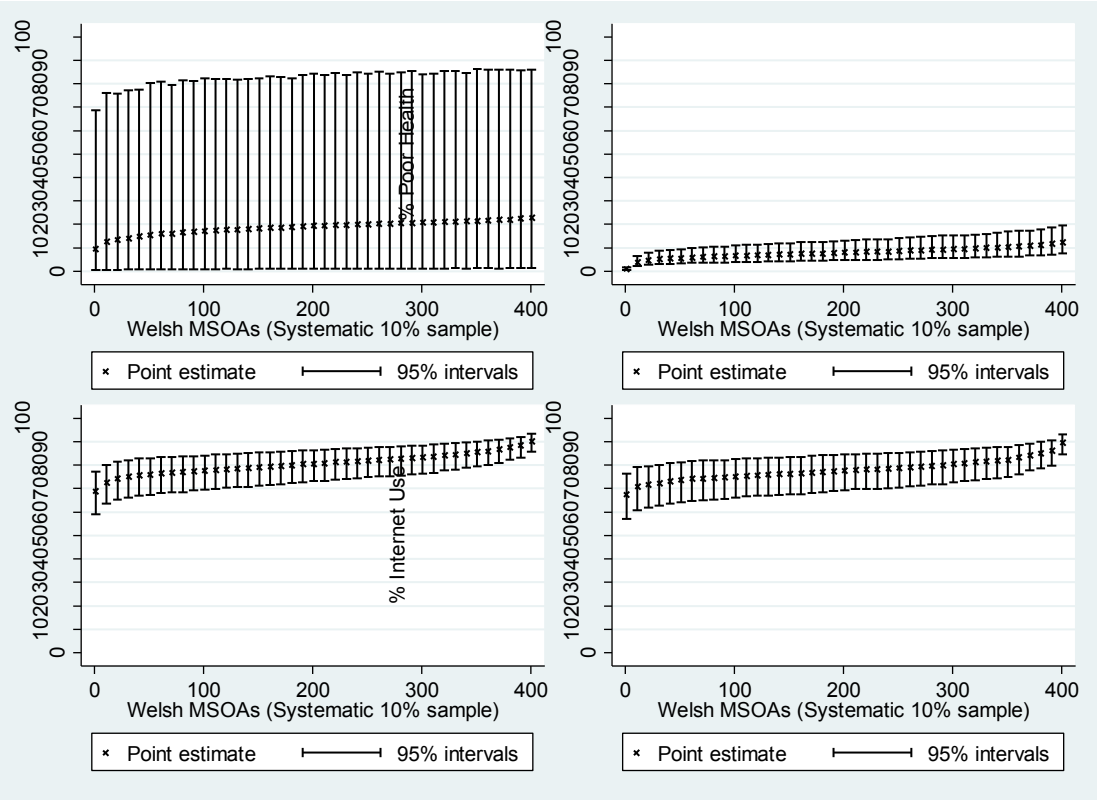
Table 7: Standard deviation of residual level 2 error for calculation of credible intervals in work strand 1

	Level 2 Units	
	LSOA	**MSOA**
Welsh Speaker	1.83	1.56
Poor Health	0.33	0.28
Internet Access	0.25	0.22
Internet Use	0.38	0.23

As would be expected, the resulting intervals are somewhat wider for the LSOA small area estimates compared with those at MSOA level. The most striking finding in Table 7 however is the markedly larger standard deviation on the residual level two error terms relating to the Welsh speaker models, resulting in far wider intervals for this outcome variable. This is to be expected given the known spatial distribution of Welsh speaking across Welsh regions and the exclusion of this factor from these individual-only specifications in work strand one. Within the context of this project as a whole the use of individual-level factors only for the estimation of this outcome in work strand one can to be considered to be a deliberately naïve omission of area-level factors that offers a foundation for comparison with the multi-level IPF in work strand two. The question for the multi-level IPF of these Welsh speaking estimates then is less *whether* the incorporation of area-level factors improves the estimates but rather by *how much* when compared to these deliberately naïve single-level estimates.

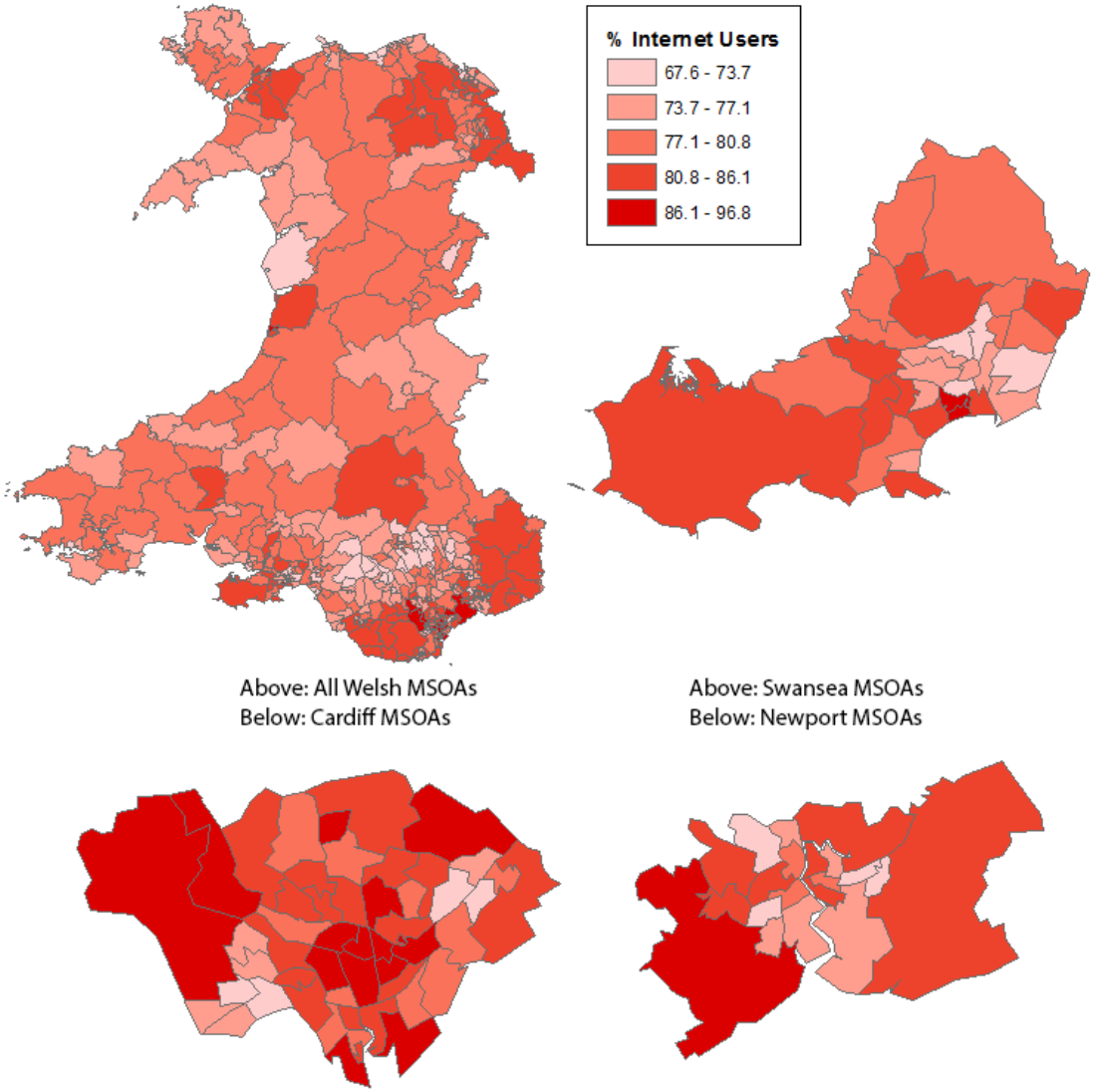
Figure 3 presents visual summaries of each set of MSOA small area estimates across a systematic ten per cent sample of their full ranges in which every 41st MSOA (i.e. 10% of the 410 Welsh MSOAs) is selected for inclusion when the MSOAs are sorted according to their estimated values on the outcome. Equivalent LSOA estimates were also created but are not shown here. The validation exercise in strand three focuses on the acceptability of these estimates alongside those produced in work strand two (i.e. estimates also incorporating area-level factors).

Figure 3: Strand one individual-level IPF small area estimates at MSOA level



In order to give a sense of the spatial detail of the resulting small area estimates, Figure 4 maps the percentage of adults who are internet users in Welsh MSOAs. Figure 4 shows results for all MSOAs across Wales (top left) as well as across Swansea (top right), Cardiff (bottom left) and Newport (bottom right). The key is based on the patterns seen at the national level but is valid for all four maps.

Figure 4: MSOA estimates of internet use across Wales



Work strand two: Multi-level IPF incorporating area-level and individual-level constraints

It is commonly the case that IPF small area estimation incorporates variables only at the individual (or household) level. One way to seek to add local specificity to small area estimates is to incorporate additional information about the area-level context alongside individual or household level constraints. One could, for example, restrict the selection of survey cases to be used in the IPF to only those survey cases falling in the same broader region or the same geodemographic classification type as the target small area (Smith et al. 2009; Birkin and Clarke 2012). Intuitively one can imagine how the tailored selection of survey cases for the different target small areas might enhance the accuracy of their small area estimation.

There is a balance between enhancing the specificity of the survey cases selected and reducing the total sample size available to fit from, as reducing the sample size can make it more difficult for the IPF reweighting to reflect the target small area's characteristics across the constraints that have been chosen. Amongst the relatively few tests which have taken place on this issue there is disagreement in findings: Anderson (2012) suggests benefits to tailored selection of cases whilst Williamson (2013) in contrast finds that the more effective strategy is usually to maximise the number of records. This may in part be a reflection of the differing spatial microsimulation methodologies used by these authors. Anderson's tests are based on an IPF approach as used in the current project. Williamson's tests, in contrast, are based on a combinatorial optimisation approach that selects the optimal subset of survey cases that best matches the small area totals across the constraints. Larger sample sizes to enhance flexibility in case selection may well be more important in combinatorial optimisation than in an IPF approach.

This second work package explores the benefits of incorporating such area-level factors in the IPF and assesses these impacts in relation to the MSOA estimates for Welsh speaking and poor health. This is of particular relevance to outcomes that show distinct variations according to area-level factors (e.g. the regional distribution of Welsh speakers) as it would not otherwise be possible to incorporate these area-level variations into the IPF process. Where area-level constraints are important for explaining the small area spatial variation in outcome variables, their inclusion is expected both to improve the accuracy of the small area point estimates and, given that a greater share of the area-level variance in the underlying multilevel regression model will be able to be accounted for, to reduce the width of the credible intervals around those point estimates.

Testing four alternative area-level constraints

This second work strand explores these issues by examining the impact of four alternative area-based strategies to the tailoring of the selection of survey cases prior to the IPF estimation. This is possible because the National Survey data includes the LSOA in which each survey respondent lives:

1. *Geodemographic type*: the Output Area Classification (OAC) 2011 uses cluster analysis across multiple socio-economic and demographic variables in

the Census 2011 to create a geodemographic classification of qualitatively distinct area types at the spatially detailed Output Area (OA) level. The base National Survey used in this project contains LSOA geocodes, a slightly larger geography into which several OAs nest. Therefore, each Welsh LSOA is placed within one of the eight OAC supergroups according to the supergroup in which the majority of its OA population lives. This LSOA level geodemographic categorisation is merged into the National Survey data and the count of individuals inside each LSOA supergroup can then be used as an additional constraint within the MSOA level IPF. We refer to this as MOAC – the Middle Layer Output Area Classification[11];

2. *Region*: a second area-based strategy to tailor case selection divides Wales into broader regions and survey cases are retained for the IPF only where they fall within the same broader region as the target small area. It is important to note that these broader regions are **not** designed primarily based on physical or territorial position in Wales but instead according to the patterns of the outcome variables seen across them, given that the aim is to draw survey cases for the IPF that are 'like' the populations of the target small areas. The design of the broader regions are therefore specific to the Welsh speaking and poor health outcome variables as shown in Figures 13 and 14 in the Appendix;

3. *Rural/Urban classification*: the Department for Environment, Food and Rural Affairs (DEFRA) 2011 Urban-Rural Classification provides an MSOA level classification of six types of rurality and urbanity. Each National Survey case is given the rural/urban grouping of the MSOA in which they live and only those survey cases within the same rural/urban group as the target small area are retained for the IPF of that small area;

4. *Welsh Index of Multiple Deprivation 2014 (WIMD)*: the WIMD provides a rich multi-dimensional measure of poverty for Welsh LSOAs and National Survey cases are placed within their respective LSOA WIMD quintile (so that, for example, the most deprived quintile contains the 20% most deprived LSOAs according to the WIMD 2014). The count of individuals inside each LSOA WIMD quintile can then be used as an additional constraint within the MSOA level IPF.

Once the impact of these four separate area-level constraints has been evaluated separately an additional two-way area selection is incorporated based on the two best performing area-level factors from the four detailed above: for the Welsh speaking outcome the best area factors are found to be region and geodemographic type whilst for the poor health outcome the best area factors are found to be region and WIMD. Two-way area classifications such as these demand relatively large survey samples to be viable and are rarely explored in the literature, although they follow logically from the use of single area classifications that have been explored in previous studies. In this two-way specification both area factors are incorporated into

[11] noting that it does not place each MSOA area into one group but rather reflects a count of the MSOA population in each supergroup at the constituent LSOA level.

the IPF as separate constraints with the result that cases are selected for the IPF only where the case falls in the same category as the target small area on *both* of these area-level variables. This adds further contextual specificity to the case selection but does reduce the number of survey cases retained for the IPF and, therefore, of the flexibility that the IPF enjoys in the reweighting process. We considered the National Survey sample size to be adequate to explore this two-way area-level combination.

Selecting the optimal IPF specification

Table 8 shows the impact of the area-level selection on the underlying model power (as measured by pseudo R-squared) in the National Survey. For each outcome variable, Table 8 firstly reproduces the model power from the optimal regression model with only individual-level factors from work strand one, as reported in Tables 5 and 6 above. For each model each of the various area-level selections is then carried out in turn alongside this base set of optimal individual-level factors. Table 8 shows three outcomes: Welsh speaking and poor health are present as the small area estimates of focus in work strand two, but findings for internet use are also shown due to interest in estimating this variable down to small area level.

A key finding from Table 8 is the need to think carefully about the source of the variation in different target outcome variables prior to any small area estimation work. This can be examined by preparing the optimal regression model of each outcome in the survey using only individual-level explanatory factors and then by comparing the effects of each set of dummy variables relating to each of these area-level variables once these are added as additional explanatory factors into that regression model. For Welsh speaking the addition of the area selection to these regression models in all cases improves the ability of the model to account for the variation in Welsh speaking. More specifically, however, it is the explanation of the *spatial* variation of Welsh speakers that is particularly helpful, with the addition of the bespoke designed region variable improving the model power from 9.6 per cent up to 31.0 per cent. The most predictively powerful combination of two area selections in the underlying regression modelling for Welsh speakers is region and geodemographic type. This two-way area categorisation further improves the model power slightly to 31.8 per cent and, as shown in Table 10 below, does improve slightly the quality of the external validation of the resulting small area estimates. In contrast, for both poor health and internet use there is no evidence that any of these area-level factors enhances the predictive power of the models over and above that given from the individual-level factors alone.

Table 8: Impact of area selection on model power within the single-level regressions

		Welsh speaking	Poor health	Internet Use
Strand One	**Individual factors only**	9.6%	40.0%	41.0%
Strand Two: Individual + Area selection	**Plus MOAC**	18.8%	40.0%	41.0%
	Plus region	31.0%	40.1%	41.0%
	Plus rural/urban	20.4%	40.0%	41.0%
	Plus WIMD	12.1%	40.0%	41.0%
	Plus two-way area categorisation	31.8%	40.1%	41.0%

With this incorporation of area selection meaning that the model is better able to account for the variation in Welsh speakers, one would expect the resulting IPF estimates to be more accurate at the small area level. The external validation process in work strand three will focus on examining this question. Additionally, however, given the methodology used to calculate the credible intervals around the point estimates one would also expect their incorporation to better account for the variance at the area level within the multilevel regression structure and, as a consequence, to narrow the width of the credible intervals. Table 9 shows that a considerable reduction in the credible intervals around the small area estimates of Welsh speakers does indeed result from the incorporation of the region variable and, to a slightly greater extent, also the two-way region-geodemographic type factors. Interestingly, although Table 8 shows that the addition of area selection makes virtually no difference to the model power when predicting poor health, Table 9 highlights that the incorporation of either the sole region factor or the two-way region-WIMD factors would result in narrower credible intervals.

Table 9: Impact of area selection on standard deviation of residual level two variance within multilevel regressions (with survey individuals nested within MSOAs)

		Welsh speaking	Poor health
Strand One	**Individual factors only**	1.56	0.28
Strand Two: Individual factors + area selection	**Plus MOAC**	1.26	0.26
	Plus region	0.66	0.20
	Plus rural/urban	1.25	0.26
	Plus WIMD	1.51	0.27
	Plus two-way area categorisation	0.60	0.20

Reflecting back on Tables 8 and 9, the optimal IPF specification for the Welsh speaking outcome is considered to be its core set of individual-level constraints identified in work strand one (and as shown in Table 5) plus the two-way region-geodemographic type area factors. For the poor health outcome the optimal IPF specification is considered to be its core set of individual-level factors identified in work strand one (again as shown in Table 5) plus the region area-level selection due

to its ability to reduce the width of the resulting credible intervals. For the poor health outcome the best performing two-way area selection (region and WIMD) is not chosen given that it does not reduce the residual level 2 error variance further but does further reduce the number of survey cases available to the IPF for the reweighting. A final consideration is the need for sensitivity to potential over-fitting through the incorporation of too many constraints and/or constraints that are too sparsely populated such that the IPF struggles to populate every cross-classified combination of the constraint factors. The addition of (especially two-way) area factors significantly affects the number of cases available for the IPF reweighting of any target small area and hence increases potential convergence problems and reliance (i.e. higher weights) on retained cases. However, analyses of the size of the resulting weights (i.e. their distribution and the reliance on cases with large weights) and the stability of the resulting estimates across different IPF specifications do not suggest concerns with these survey data.

Visualising the small area estimates

Figure 5 presents the resulting MSOA level IPF point estimates for the percentage of adults in poor health (left, in green) and the percentage of adults who speak Welsh (right, in purple). A concentration of higher levels of adults in poor health is estimated across the Welsh Valleys whilst Welsh speakers are estimated to concentrate down the west side of the country and particularly within Gwynedd, Anglesey, Ceredigion and Carmarthenshire. The task of work strand three is to assess the degree to which these estimate patterns are accurate at this detailed MSOA geography.

Figure 5: IPF estimates of the percentage of adults in poor health

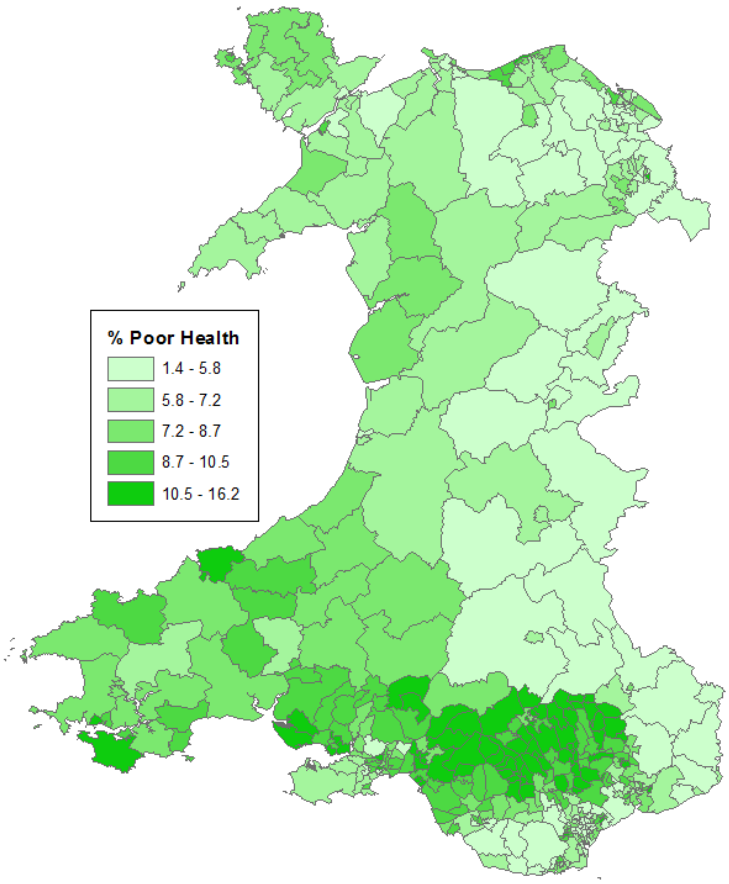


Figure 6: IPF estimates of the percentage of adults speaking Welsh

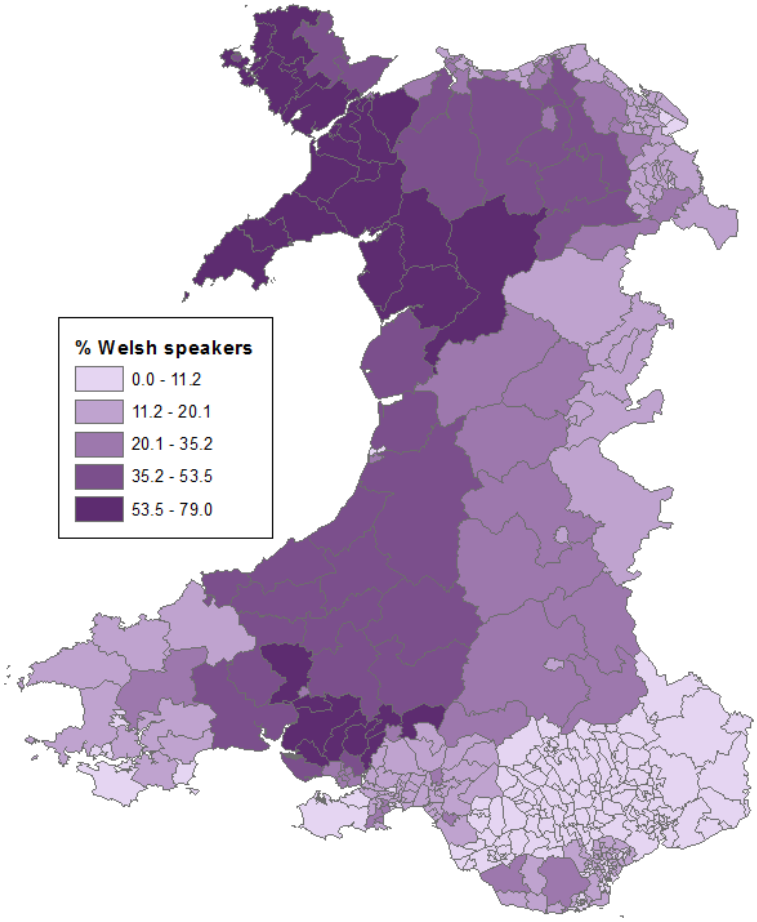
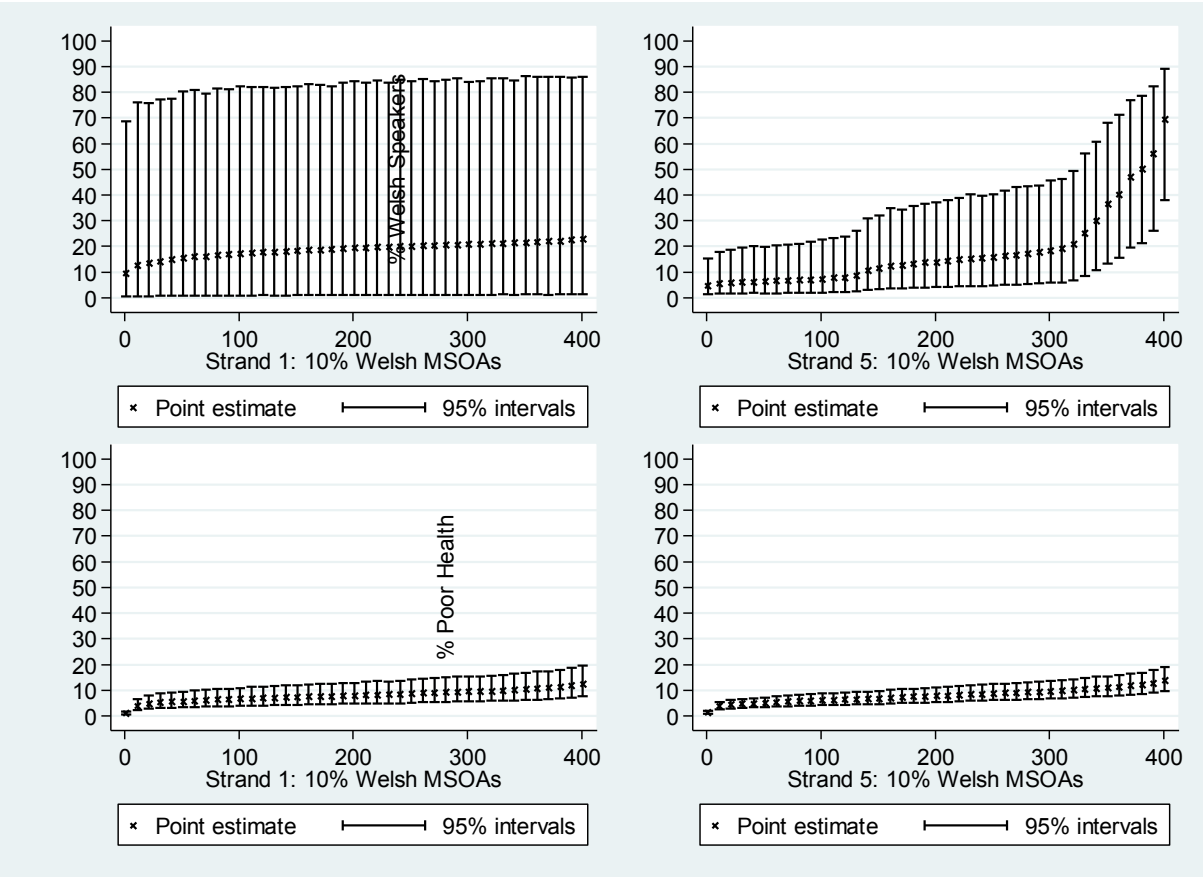


Figure 7 presents caterpillar plots of a systematic ten per cent sample of the full range of the small area estimates for the percentage of MSOA adults who speak Welsh (on the top row) and the percentage of MSOA adults in poor health (bottom row) based on the optimal IPF specification as described above. In order to more easily compare the impact on the final small area estimates of incorporating the area factors on each row the left chart repeats the MSOA estimates from strand one (when only individual constraints are used,) whilst the right chart shows the MSOA estimates once the selected area factors are added to those selected individual constraints (region and geodemographic types for Welsh speakers, region for poor health).

As noted above, the width of the credible intervals can be seen to narrow somewhat for the poor health estimates although the point estimates themselves look relatively unchanged. The incorporation of the region and geodemographic type constraints causes dramatic changes to the estimates of Welsh speaking however. Most obvious is a marked reduction in the width of the credible intervals, although there remains a reasonable amount of between-MSOA variation still unaccounted for, as seen in the continued relatively wide ranges of these intervals. It will be noted that the credible intervals tend both to narrow and to be asymmetrical as point estimates approach the 0% or 100% values, a result of the greater confidence in results which deviate towards these extremes when calculating intervals around percentages. Also apparent is the greater ability of the IPF to reflect the spatial variability in the percentage of adults speaking Welsh across these MSOAs once these area

constraints are incorporated. This can be seen most clearly towards the higher estimates in the top right chart of Figure 7 where the IPF is now more able to stretch the higher point estimates to far higher levels than those estimated in the top left chart when only individual level constraints are used.

Figure 7: Comparison of the IPF estimates of Welsh speaking (top row) and poor health (bottom row) from the individual-only IPF (left-hand side) and the individual-plus-area IPF (right-hand side)



Work strand three: Validating the estimates

A key part of any small area estimation exercise is the need to validate the resulting estimates as fully as possible. Validation can be separated into internal validation – concerning statistics of constraint fit and model power internal to the IPF – and external validation – concerning the extent to which the resulting small area estimates align with other known data of equivalent outcome variables that are thought to be 'true' (though recognising that often they are not necessarily 'true': for example, direct survey estimates always have a margin of error). Understandably, this process of external validation is of critical interest because of the desire to know how much confidence one can place in the accuracy of the small area estimates. Yet external validation is typically also difficult to execute given that it is the lack of such data at the small area level that is usually the motivation for needing to use small area estimation in the first instance.

Internal validation of the estimates

Internal validation relates checking the adequacy of the estimation model based on characteristics such as the predictive power of the regression models underlying the IPF. In the current project, this check suggests, for example, that the small area estimates of Welsh speaking from the strand one IPF with only individual level covariates ought to be treated with significant caution (because the regression model has a pseudo-R^2 of 9.6%). It also suggests that once area selection is incorporated in strand two, the estimates of Welsh speaking are markedly more accurate (given that the regression model including this area selection has a pseudo R^2 of 31.8%).

Given that an IPF approach seeks to reweight survey cases such that in the aggregate they fit the totals of a set of selected constraint characteristics, internal validation of IPF can be assessed by seeing how closely the weighted sums of these constraint characteristics in the survey compare to the 'true' constraint totals for the small areas.

Tables 16 to 19 in the Appendix present the results of these internal validation checks for work strands one and two of the project. Firstly, the mean standardised absolute error is calculated for each small area by taking the difference between the IPF's weighted constraint total and the actual small area Census total and then expressing that difference as a percentage of the actual Census total. The average of this value across all the small areas is then reported as a summary measure of constraint fit. A second fit measure shows the percentage of the small areas (either LSOA or MSOA depending on the work strand) that have weighted survey constraint totals within 20 per cent of the actual Census totals.

In the unshaded cells Tables 19 to 21 shows these results for the actual constraints used within each IPF specification. The IPF achieves excellent fit to the Census totals across the constrained variables, as is typically the case for an IPF approach. A more demanding challenge is to question the extent to which the IPF is also able to achieve good fit across factors *not* included as constraints. This is based on the logic that one is seeking through the IPF to create synthetic micro-populations that mimic the actual small area populations across the multi-dimensional range of

characteristics. At the same time it should be noted that such non-constrained variables have by definition been found not to be influential to the prediction of the outcomes within the underlying regression models and hence the impact of weaker fit on such factors to the accuracy of the final estimates is arguably less important. Within the grey shaded cells Tables 19 to 21 show an illustrative selection of the fit statistics across non-constrained factors and these are indeed less well matched with the Census counts compared to the fitted constraints.

External validation of Welsh speaking and poor health IPF estimates

The current project's aims are around methodological learning at least as much as the generation of new small area estimates. As such, in order to support the most robust external validation exercise possible, the project produces small area estimates for two outcome variables that are already available at small area level from Census 2011 (the percentage of adults speaking Welsh and in poor health respectively). External validation can therefore be carried out by directly comparing the IPF estimates against the known 'true' values at those same small area scales as taken from Census 2011. In contrast, no robust small area data is available from other sources for two other outcomes covered by the project, internet use and internet access. For these outcome variables the external validation is restricted to a comparison of the IPF small area estimates aggregated up to local authority level and compared against weighted direct survey estimates from the National Survey 2013-14 across those local authorities.

The Welsh speaking and poor health outcomes were included in both of the first two project work strands:

- In work strand one, both outcomes were estimated at LSOA level and MSOA level using only individual-level constraints;
- In work strand two, both outcomes were estimated at MSOA level with selected area constraints added to the pre-selected set of individual constraints.

Looking firstly at the LSOA estimates produced in work strand one, using only individual factors, it will be remembered that these underlying binary logistic regression models are far better able to account for the variation in poor health than the variation in Welsh speaking. One would expect the external validation to reflect this differential model power. Indeed, the gap between the two widens in the external validation exercise. At LSOA level when we compare the IPF estimated percentages at this small scale to the 'true' percentages taken from the Census 2011 the Pearson's correlation coefficient between the two sets of values is 0.93 for the percentage of adults in poor health at LSOA level but is only 0.10 for the percentage of adults speaking Welsh. Individual factors alone are able to acceptably estimate the poor health outcome but not the Welsh speaking outcome.

Turning next to the MSOA IPF estimates for these two outcomes from work strand two, i.e. once area constraints are incorporated, Table 10 shows similar correlation coefficients between their respective IPF estimates and the 'true' MSOA percentages taken from the Census 2011. Shaded in blue are the two IPF specifications that the

model testing suggested to be the best specification for the IPF. Although here we can compare the external validation performance of this pre-selected 'best' specification with those other specifications, under usual circumstances 'true' values would not be known at the small area level and we would be reliant on the model testing phase to select the single specification to use in producing the small area estimates. These results support the view that it is sensible to require model power to be adequate (reaching for example beyond a minimum level of 30% in terms of an R^2 value) in order to generate small area estimates that could be expected to validate well externally.

The final row of Table 10 shows additional MSOA estimates provided by Ipsos MORI using an area-level regression modelling approach within a separate project (Ipsos MORI, 2015, see Annex A for the full report). Its inclusion enables us to compare the performance of the IPF spatial microsimulation approach used in this project with an alternative small area estimation approach based on regression models using area-level factors only.

As expected, for the percentage of adults speaking Welsh the external validation performance is poor when fitting only individual-level constraints in the IPF but improves dramatically once area-level constraints are incorporated. From those model power values in Table 8 it is the region and combined region-MOAC specifications that were expected to be the best specifications for the IPF and this is indeed the case: the correlation with the 'true' Census percentage is 0.91 when the regional constraint is added and rises further to 0.93 when the combined region-MOAC constraints are added. This is the specification that was selected as optimal in advance. The Ipsos MORI estimates also achieve a correlation of 0.93 on the Welsh speaking outcome.

For the percentage of adults in poor health the IPF specification with only individual level constraints performs well in terms of its underlying model power (see Table 5) and little to no additional model power is brought by incorporating area factors. A similar picture is seen in terms of the external validation in Table 10 where there is relatively little difference between the external validation performance of these various sets of IPF estimates and where all of the IPF specifications slightly outperform those from the Ipsos MORI ecological modelling approach. Interestingly, although these IPF validation results are all relatively similar across the different IPF specifications it is not the specification selected in advance on the basis of the underlying model testing that emerges as the best performing in terms of the external validation. The implication is that whilst the underlying model is a strong guide to the *general* performance of the IPF estimation it is not possible to assume *specific* performance results in terms of the external validation (even if these external validation results across the IPF specifications are close to one another).

Table 10: Correlations between MSOA small area estimates and 'true' Census 2011 MSOA values

% adults speaking Welsh in the MSOA			% adults in poor health in the MSOA		
Work strand One	Individual only	0.08	Work strand One	Individual only	0.93
Work Strand Two	Plus MOAC	0.65	Work Strand Two	Plus MOAC	0.91
	Plus rural/urban	0.67		Plus rural/urban	0.94
	Plus region	0.91		Plus region	0.92
	Plus WIMD	0.39		Plus WIMD	0.93
	Plus region-MOAC	0.93		Plus region-WIMD	0.91
Ipsos MORI estimates		0.93	Ipsos MORI estimates		0.89

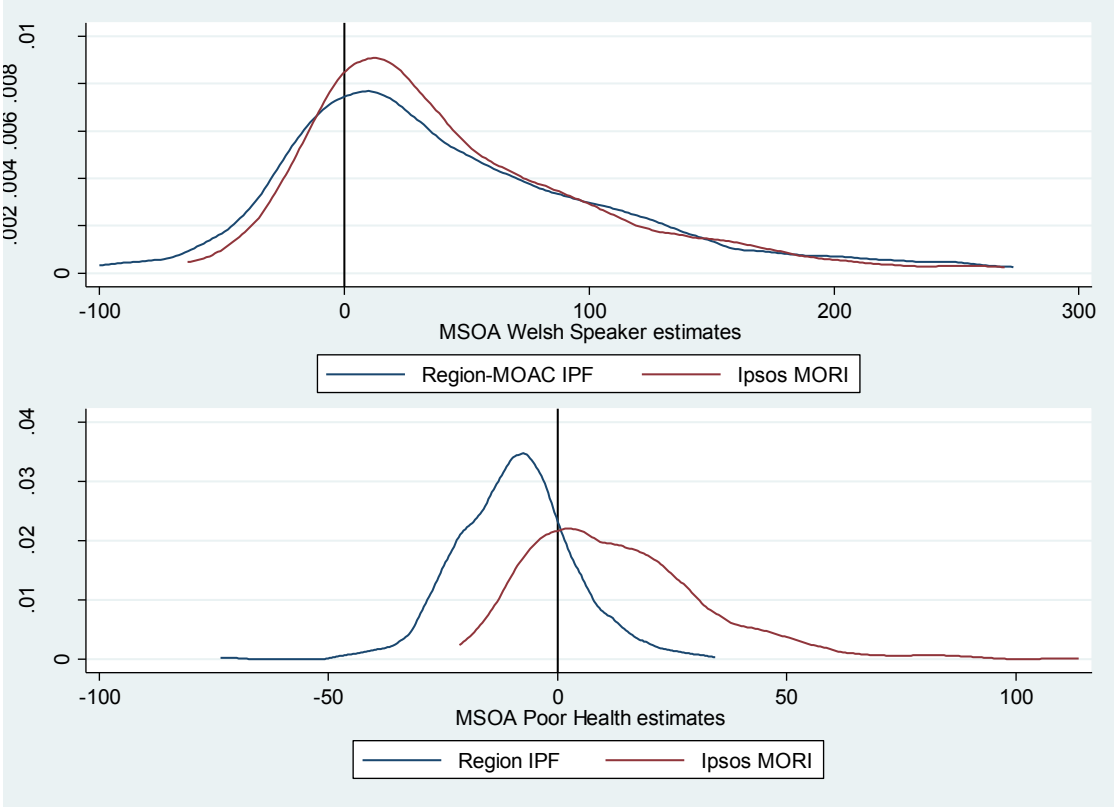
Although the correlation coefficients in Table 10 offer concise summary measures of the external validation performance, Figure 8 focuses further on the comparison of the selected 'best' IPF specifications[12] with the Ipsos MORI estimates. Figure 8 shows kernel density plots of the standardised error of these two sets of estimates across all 410 MSOAs, where standardised error refers to the difference between each MSOA's IPF estimated percentage and its 'true' Census percentage expressed as a percentage of the MSOA's 'true' Census value. A value of zero therefore implies the IPF estimate matches the Census value exactly, positive values imply that the estimates are higher than the Census percentages, and negative values imply that the estimates are lower than the Census percentages.

For the MSOA percentage of adults speaking Welsh (top chart) the two distributions are extremely close to one another, particularly above zero in terms of the degree of over-estimation. The Ipsos MORI estimates have a slightly higher number of MSOAs concentrated around the zero value with very well matched estimates whilst the IPF estimates have a small number of MSOAs with larger under-estimates compared to the Ipsos MORI under-estimates.

For the MSOA percentage of adults in poor health (bottom chart) the two sets of estimates are more diverse. For the IPF estimates there is a tendency towards slight under-estimation and with a small number of MSOAs with relatively large negative values. For the Ipsos MORI estimates the errors are instead towards slight over-estimates, with the Ipsos MORI mean being 10.2 compared to the IPF mean being -8.9. The Ipsos MORI estimates show a wider distribution than the IPF estimates (standard deviation of 19.9 compared to 12.8 in the IPF) and have a larger and longer tail of relatively large positive values of markedly over-estimated MSOA values compared to the tail of under-estimated MSOAs seen in the negative tail of the IPF distribution.

[12] Which are, to recap, individual constraints plus region and geodemographic type for Welsh speaking and individual constraints plus region for poor health.

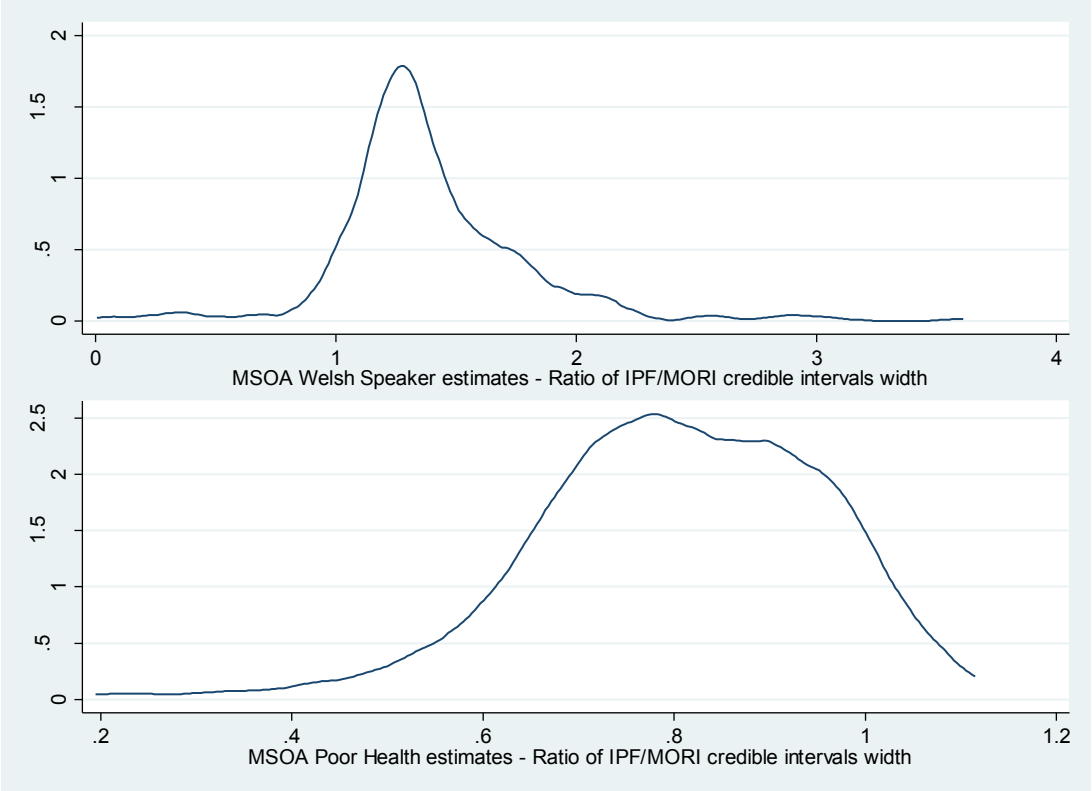
Figure 8: Standardised error of best IPF specification vs best Ipsos MORI specification



An additional area of interest is to explore the nature of the credible intervals both in terms of their width and in terms of their performance in containing the 'true' Census values. Figure 9 shows ratios between the width of the credible intervals in the same 'best' IPF specifications and the Ipsos MORI ecological modelling approach. The ratio is expressed as IPF/Ipsos MORI such that a value of one implies that both sets of credible intervals are the same width, positive values imply that the IPF intervals are wider than the Ipsos MORI intervals (with a value of 2 being twice as large for example), and values less than 1 imply that the IPF intervals are narrower than the Ipsos MORI intervals (0.5 being half as wide for example).

For Welsh speakers (top chart) the IPF intervals are generally wider than the Ipsos MORI intervals, with a median ratio of 1.3 across all MSOAs. For the poor health credible intervals (lower chart) the trend is in the opposite direction with the IPF intervals generally narrower than the Ipsos MORI intervals, with a median ratio of 0.81 across all MSOAs. As shown in Table 8 above, this reflects differences in the extent to which the variation in these two outcomes is accounted for more by individual (poor health) or area (Welsh speaking) factors and the corresponding extent to which these two alternative small area estimation implementations are based on factors at the area level (Ipsos MORI exclusively) or across both levels (individual level predominantly plus some incorporation of area factors in the IPF).

Figure 9: Comparison of the width of the credible intervals of best IPF specification vs best Ipsos MORI specification



Ultimately, the function of the credible intervals is to offer a range within which one can be statistically confident to the 95% level that the 'true' underlying population values fall. Table 11 examines the extent to which both these sets of intervals satisfy this core function. For both outcomes the vast majority of the 'true' MSOA Census percentages fall within the respective credible intervals from both the IPF and Ipsos MORI estimates (85% of the MSOA estimates of Welsh speaking and 92% of the poor health estimates). Overall the IPF intervals perform somewhat better than the Ipsos MORI intervals. For Welsh speakers 92.2% of 'true' Census percentages fall within the IPF estimated MSOA intervals compared to 89.5% within the Ipsos MORI estimated intervals, although those IPF intervals are generally slightly wider as seen above. For poor health 96.3% of 'true' Census percentages fall within the IPF estimated credible intervals compared to 95.4% within the Ipsos MORI estimated intervals, despite the IPF intervals also being narrower on average.

Table 11: 'True' MSOA values falling outside the estimated credible intervals

% Welsh Speaker MSOA estimates						% Poor Health MSOA estimates				
		MORI						MORI		
		In	Out	Total				In	Out	Total
IPF	In	350 85.4%	28 6.8%	378 92.2%		IPF	In	378 92.2%	17 4.1%	395 96.3%
	Out	17 4.1%	15 3.7%	32 7.8%			Out	13 3.2%	2 0.5%	15 3.7%
	Total	367 89.5%	43 10.5%	410 100%			Total	391 95.4%	19 4.6%	410 100%

External validation of the internet use and internet access IPF estimates

Work strand one of the project produced estimates of internet use and internet access at both LSOA and MSOA scales based on an IPF specification with individual-level constraints only. Based on the predictive power of the underlying regression models, Table 8 above suggests that it is not expected for there to be much, if any, improvement to the estimates by incorporating additional area-level factors to this IPF specification. Given that known values of internet use and internet access do not exist at these small scales the external validation takes place by aggregating their IPF small area estimates to the local authority level and comparing them at this higher scale against the weighted survey estimates derived directly from the National Survey 2013-14. To do so the mean survey estimate is taken as 'true', although it is acknowledged that this ignores the reality that there are confidence intervals around those mean survey estimates.

Figure 10 shows scatter plots of these comparisons. The LSOA estimates are shown across the top row of Figure 9 whilst the MSOA estimates are shown along the bottom row; the internet use estimates are shown down the left side of Figure 10 whilst the internet access estimates are shown down the right side. A line of equality is added to each chart. The patterns seen are stable across these two internet outcome variables (a reflection of the strong correlation between them) as well across these two small area scales. There is a general tendency towards slight over-estimation from the aggregated IPF estimates compared to the direct National Survey estimates and no evidence of any local authorities being outliers in terms of markedly poorer fit to their National Survey equivalents.

Figure 10: Local authority level external validation of internet use and internet access estimates

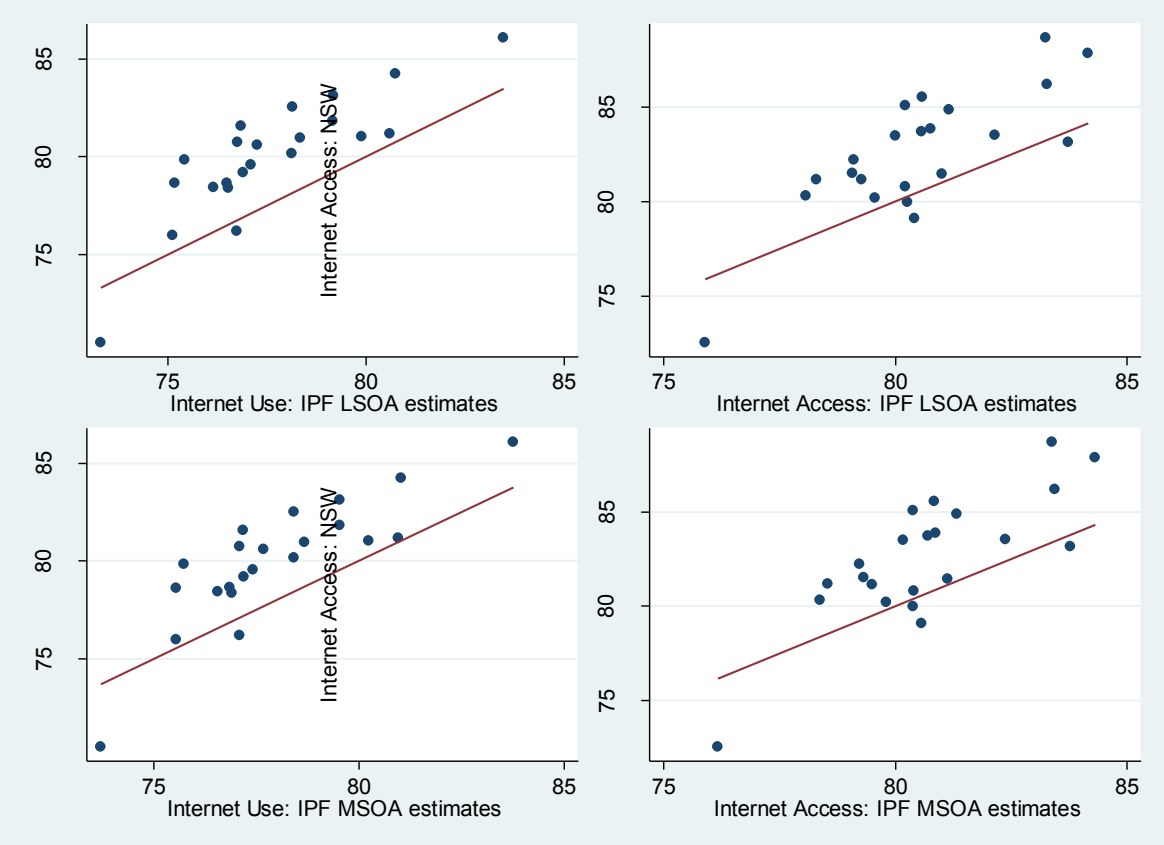


Table 12 formalises these visual impressions through correlation coefficients between the aggregated LSOA and aggregated MSOA IPF estimates against the National Survey estimates for each of Wales' local authorities. All correlations are strong with the internet access coefficients slightly stronger than those for internet use. Although somewhat lower than equivalent correlations from other studies these values are broadly in line with those other findings (c.f. Anderson, 2007; Tanton et al., 2014).

There is no difference in the external validation statistics of the same outcome between the MSOA and LSOA scales. In terms of whether these external validation statistics ought to be considered strong enough, it is noted that decisions around the acceptably of external validation statistics ultimately depend in part upon subjective decisions around the use to which the small area estimates will be put and the extent to which analysts and policy-makers demand likely accuracy at the small scales for those purposes.

Table 12: Correlations between aggregated IPF estimates of internet use and internet access and direct National Survey local authority estimates

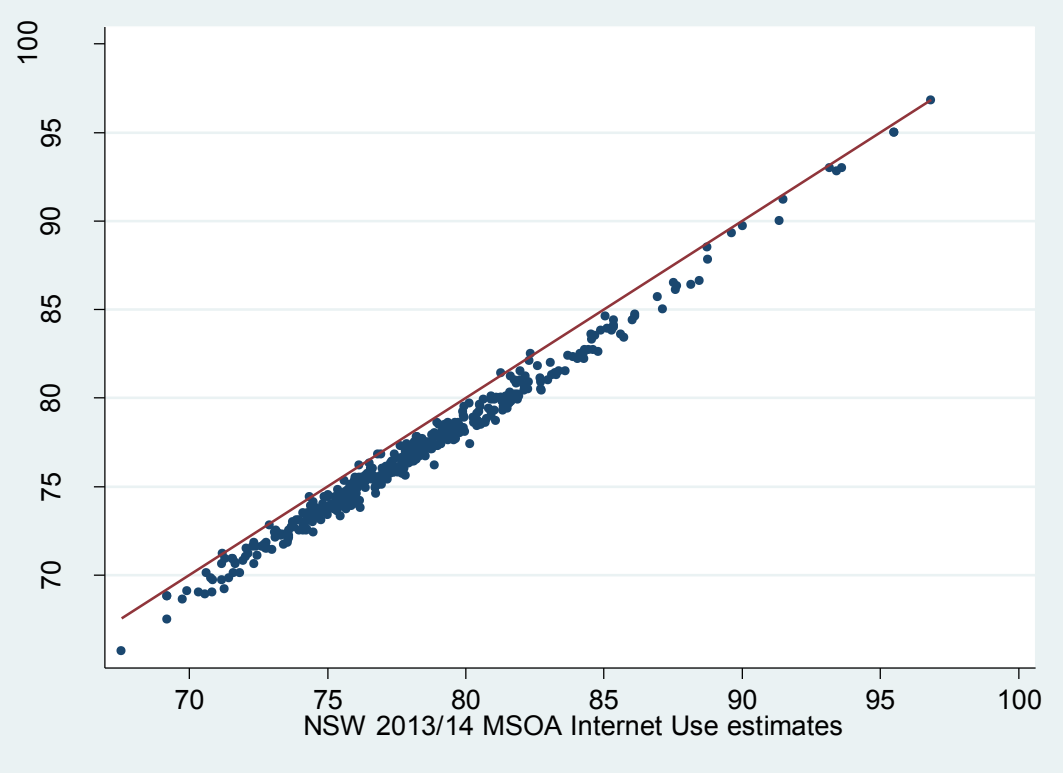
	Internet Use	Internet Access
MSOA IPF estimates	0.80	0.84
LSOA IPF estimates	0.80	0.84

Comparing IPF estimates of internet use between National Survey 2012-13 and National Survey 2013-14

One area of additional interest within the external validation is to compare the MSOA estimates for the percentage of adults using the internet produced in this current project using the National Survey for Wales 2013-14 against equivalent MSOA estimates produced during a previous small area estimation project using the same IPF methodology but based on the National Survey for Wales 2012-13 (Whitworth et al., 2015). Figure 11 below shows a scatter plot of the two sets of IPF point estimates with the 2012-13 survey data on the vertical axis and the 2013-14 survey data on the horizontal axis. A line of equality between the two sets of estimates is also shown.

The two sets of estimates show a strong fit to one another with stability across these two years. More formally, the Pearson's correlation coefficient between the two sets of MSOA estimates is 0.99. A slight shift rightwards in the markers is evident, implying that the IPF estimates are suggesting a small overall increase in the percentage of adults using the internet at this small area level. If one examines the raw National Survey for Wales data then this increase is evident at the national level, with weighted direct survey estimates for personal internet use at the national level rising from 76.9% in the National Survey 2012-13 data to 80.65% in the National Survey 2013-14 data. Although the IPF therefore correctly produces somewhat higher MSOA estimates of internet use for 2013-14 compared with 2012-13 this increase does not reflect the full extent of the increase seen nationally.

Figure 11: Comparison of MSOA estimates of internet use between National Survey 2012-13 and 2013-14 data



Work strand four: Post-estimation constraining

Work strand two focuses on optimising the IPF specification *prior* to the process of small area estimation in order to optimise the quality of the resulting small area estimates. The external validation process in work strand three has shown that this is generally helpful and that for some outcome variables it can be critical to the success of the estimation process.

In contrast, work strand four explores the potential for constraining resulting small area estimates to sum to known 'true' values at higher spatial scales. The logic of such constraining approaches is intuitive: if the IPF estimated small area counts of adults affected by each outcome are summed to the local authority level, then ideally those summed counts should equal the local authority counts that are known and taken to be 'true' from either the Census 2011 or the National Survey. Where they do not then if the Census 2011 counts or direct survey counts at local authority level are taken to be 'true' – acknowledged to be something of an assumption, particularly for survey based local authority estimates – one can seek to adjust the small area level IPF estimates such that they sum to those totals. They might then be considered 'correct' at these aggregate scales, although it is important to acknowledge that such constraining does not guarantee accuracy in the small area distributions of the constrained IPF estimates.

The focus in this work strand is on the MSOA estimates of the percentage of adults in poor health (a variable with known 'true' values at the small area level from Census 2011) and the MSOA estimates of internet use (a variable with known 'true' values only at the higher local authority level via direct estimates from the base National Survey for Wales data). The MSOA internet use estimates were produced in work strand one whilst the MSOA poor health estimates used are the 'best' specification selected (those combining individual constraints with region in work strand two). The interest in terms of future estimation is inevitably on outcome variables such as internet use where the impact of any such constraining cannot be validated externally at small scale. However, the inclusion of the poor health outcome in this work strand acts both to support the construction of a spatially varying constraining approach for the internet use estimates as well as to allow external validation at small area level of the impact of such constraining.

Focusing firstly on the internet use outcome, the essence of the constraining is to take the difference between the 'true' count of internet users in each local authority (as given by survey weighted analyses of the National Survey for Wales 2013-14, ignoring the confidence intervals around the central point estimates) and the total local authority IPF count (as given by aggregating the MSOA level IPF estimated counts of internet users) and to reallocate this 'error' back across the MSOAs inside the local authority such that the adjusted MSOA IPF counts sum to the 'true' local authority counts. Three alternative constraining techniques are explored:

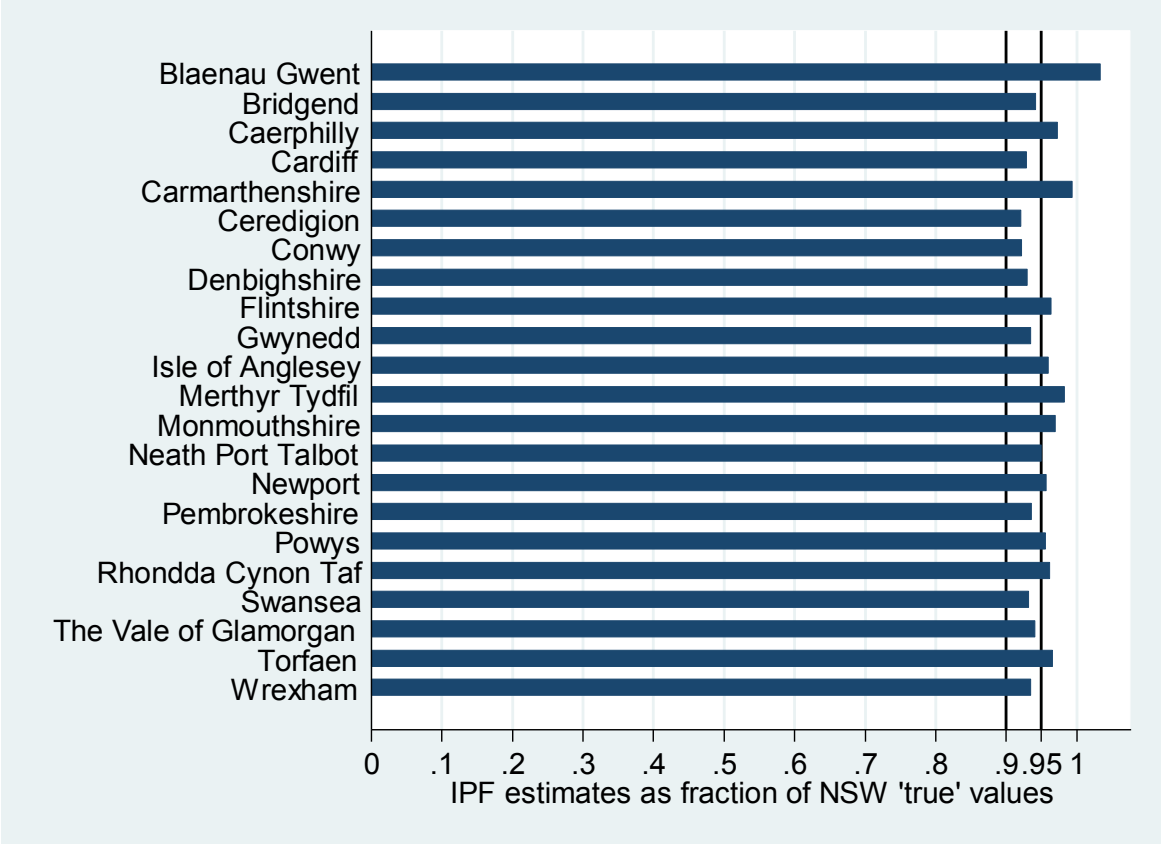
- Errors divided equally across MSOAs: the total local authority error is split equally across the constituent MSOAs so that all MSOAs inside the same

local authority take the same absolute level of adjustment to their estimated counts;

- Errors divided in proportion to the estimated count of internet users in each MSOA: the IPF estimated counts of internet users will vary across MSOAs inside any local authority and it seems reasonable to suggest that the extent of error in any MSOA estimate will vary in proportion to the size of that base count;
- Errors divided according to sub-local authority knowledge about the spatial variation in errors of a separate variable known at small scale: a third possibility is to calculate the errors between estimated and 'true' values at MSOA level for a variable whose 'true' value is known at small scale (i.e. the number of adults in poor health) and to use those small scale 'errors' as weights to fractionally apportion the total local authority error seen for an outcome that is not known at small scale (i.e. internet use) across its IPF estimated counts of internet users in constituent MSOAs. Such an approach potentially enables the constraining process to incorporate knowledge of the spatial patterning in sub-local authority error but does assume that it is reasonable to apply those patterns from one variable (poor health) onto another (internet use).

Figure 12 provides the context for these analyses. For each local authority it shows the fraction between the local authority sums of the IPF MSOA estimates of adults using the internet and the local authority count of such adults as estimated directly from the National Survey 2013-14. As implied by the external validation of this outcome from work strand three, the fit between these two sets of local authority counts is relatively tight: all authorities show fractions over 0.9 and half show fractions above 0.95. Blaenau Gwent is the only local authority where the aggregated IPF estimates are greater than the direct local authority survey estimates. Inevitably there remains a degree of error between the two sets of estimates, with the aggregated IPF counts never summing perfectly to the 'true' National Survey counts at local authority level. It is this gap that the constraining seeks to rectify.

Figure 12: Total local authority count of IPF estimated internet users as a fraction of National Survey point estimates of internet users



All three alternative constraining approaches outlined above serve to ensure that the estimated counts of adult internet users at the MSOA level sum to the 'true' counts at local authority level as derived from the National Survey. Of the three constraining approaches examined (shown in Table 13) the first two alternatives make relatively modest adjustments to the MSOA estimates of internet use when correlated back against the original unconstrained IPF estimates. However, it is notable that the attempt to create an allocation of error based on knowledge of the sub-local authority spatial variation in errors from the poor health outcome results in a dramatically greater adjustment to the MSOA internet use estimates, as shown by the far weaker correlation between these estimates and the original unconstrained estimates.

One issue within this constraining approach is also the difficulty in handling cancellation between positive and negative errors in the poor health estimates at the MSOA level whilst also requiring the constraining to sum to local authority error seen in the internet use outcome. As shown in Figure 11, with the exception of Blaenau Gwent the aggregated IPF estimates of internet use required local authorities to 'take on' *additional* internet users in order to meet local authority survey values in this work strand. All errors were therefore converted to positive values and all MSOA estimates of internet use were uprated for these local authorities; for Blaenau Gwent the opposite is the case and MSOA estimates were universally deflated for this local authority. This is acknowledged as an assumption. Taken together, the assumptions within this spatially varying constraining approach and the size of the adjustments made by this approach as outlined in Table 13 lead us to advise against use of this spatially varying approach to post-estimation constraining.

Table 13: Impact of the constraining on the IPF estimates of the percentage of adults using the internet across Welsh MSOAs

IPF MSOA estimates	Pearson's correlation with original unconstrained MSOA estimates of the percentage of adults using the internet
Original IPF estimates constrained evenly across MSOAs within each local authority	0.92
Original IPF estimates constrained in proportion to the estimated count of internet users across MSOAs within each local authority	0.95
Original IPF estimates constrained according to the spatial patterning of error in the MSOA IPF estimates of the number of adults in poor health	0.73

Ideally one would additionally wish to assess the impact of these constraining approaches on the original estimates at the small area scale but a clear challenge is the lack 'true' figures for internet users at these scales. In order to seek to shed light on impacts at this small scale the constraining approaches were therefore applied to the IPF MSOA estimates of poor health – an outcome with known MSOA values from the Census 2011 – in order that equivalent external validation can take place at the small area MSOA scale. Naturally it is not necessarily the case that lessons from this outcome can be applied to all outcomes but these analyses do offer valuable insights. The spatially varying constraining approach is not employed for the poor health outcome, both because of the negative findings above which suggest this approach ought not to be adopted, as well as due to the circular nature of this approach for this outcome.

Table 14 shows the resulting correlations between 'true' percentage of adults in poor health as according to the Census 2011, the original set of unconstrained IPF estimates produced in work strand two and the two sets of alternatively constrained IPF estimates. To clarify, whilst Table 13 above presents these correlations for internet use at local authority level Table 14 is able to assess the impact of the constraining on the poor health estimates at the target MSOA scale. With a Pearson's correlation coefficient of 0.93 the set of MSOA IPF estimates of poor health produced in work strand two (those incorporating region at the area level) already correlate strongly with the 'true' MSOA Census percentages. Following constraining the strength of the correlation increases further still to a value of 0.96 when allocating the local authority error evenly across all MSOAs within that authority and to a value of 0.97 when allocating the local authority error in proportion to the IPF estimated count of adults in poor health within the local authority.

Table 14: Impact of the constraining on the IPF estimates of the percentage of adults in poor health across Welsh MSOAs

IPF MSOA estimates	Pearson's correlation with 'true' MSOA percentages of adults in poor health from Census 2011
Original unconstrained IPF estimates	0.93
Original IPF estimates constrained evenly across MSOAs within each local authority	0.96
Original IPF estimates constrained in proportion to the estimated count of adults in poor health across MSOAs within each local authority	0.97

Clearly such constraining approaches involve assumptions – Is the local authority survey estimate for internet users really 'true'? How can we be sure that the allocation of error matches the real sub-local authority patterning of error? Do we make some small area estimates worse even through in the aggregate we improve the estimates? However, these analyses suggest that viable constraining methods do exist that are relatively simple to implement post-estimation and that might be incorporated to slightly refine and enhance small area estimates produced from the IPF.

Work strand five: Estimating healthy lifestyles

Work strand five represents a stand-alone element of the overall project that assesses the viability of creating small area estimates of four outcomes relating to healthy lifestyles using the Welsh Health Survey data. In keeping with the emphasis on methodological learning across the project, work strand five seeks to compare the resulting IPF estimates to existing direct survey estimates for these outcomes as well as to explore the impact on estimates of using a variety of differently designed smaller base survey files.

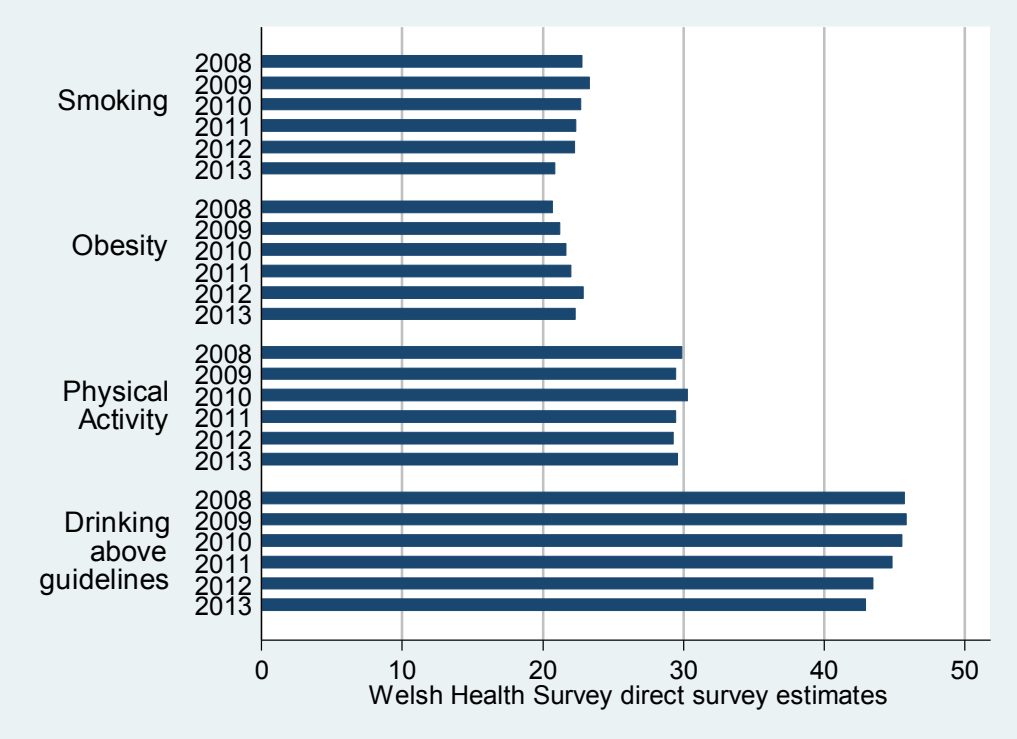
The four outcomes in focus across work strand five are the percentage of adults who:

- smoke[13];
- are obese;
- were physically active on at least five days in the previous week; and
- drank above alcohol guidelines at least once the previous week.

Figure 13 provides the national figures for these four healthy lifestyle outcomes across Wales based on the weighted direct survey estimates from Welsh Health Survey data. A little over 20 per cent of Welsh adults smoke, a similar proportion are obese, just under 30 per cent are physically active, and close to forty-five per cent are identified as drinking above guidelines. These outcomes vary relatively little across the six years of Welsh Health Survey data used and so whilst some change over time is smoothed over this is relatively small and is outweighed by the advantages of the notably larger sample size. At the same time, it is noteworthy that the results below highlight the potential of the small area to produce acceptable estimates from smaller base survey files than the six year pooled dataset used. This offers potential benefits in the ability of the IPF to rely on more recent survey data (e.g. 2011 to 2013 rather than 2008 to 2013 survey data) and, in doing so, to reduce the extent to which change over time in the outcome variables of interest is smoothed over.

[13] See Table 1 for a detailed description of these outcome variables.

Figure 13: Welsh Health Survey estimates of the percentage of Welsh adults for each outcome



IPF estimates of these four outcomes are produced at the Upper Super Output Area (USOA) scale, a larger statistical geography in Wales that is made up of several MSOAs and that nests inside local authority boundaries. The USOA IPF estimates are created using a pooled dataset of the Welsh Health Survey results from 2008 to 2013, and are created in unstandardized form as well as in standardised form based on the European Standard Population age structure as provided by the Welsh Government. The resulting IPF estimates are compared to direct survey estimates of these four outcomes that exist already at the USOA scale from survey weighted direct estimates of the same Welsh Health Survey data from 2008 to 2013.

As in previous work strands, in order to select the optimal set of IPF constraints for each outcome variable a series of binary logistic regression models are explored using model power as the guide to selection. Care is taken not to over fit across too many constraints and/or constraints that are too sparsely populated. Constraint selection operates across three phases. Firstly, models are used to identify the optimal set of individual-level factors. Secondly, the four separate area-level selection factors used in the previous work strand (geodemographic type, rural/urban classifications, region[14], WIMD quintiles) are introduced and their impacts on model power are evaluated. Finally, the two most predictively powerful area-level selection factors are then both retained and their combined effect on model power is assessed. Of these alternative specifications the optimal sets of constraints is then selected.

Table 15 below shows the final set of optimal constraints for each outcome variable following this testing process. In all cases, two area-level constraints are retained in order to maximise model power. The gains to model power are relatively marginal

[14] Based on the design of regions as used for the poor health outcome in work strand two.

compared to the incorporation of just the single most predictively powerful area-level factor. However, the very large survey sample size (n=62,269) enables these marginal gains from a more detailed constraint specification to be realised without notable cost to the model fitting procedure (see discussion in the introduction to work strand two).

The presence of limiting long-term illness as an individual-level factor is worthy of note. This variable only appears in the survey years from 2011 to 2013 and so where it is found to be a predictively powerful explanatory factor in these pooled models its inclusion results in roughly halving the sample size available for the IPF. Given the very large total sample size of the pooled WHS data, however, this is not problematic.

The two shaded rows at the bottom of Table 15 show two sets of key statistics around the small area estimation of these four healthy lifestyle outcomes. Firstly, the predictive power of the underlying binary logistic regression models used to select the optimal IPF constraints is shown. None of these models could be considered particularly powerful and this weak predictive ability does not offer the strongest foundations for the small area estimation. The variables included in these regression models – noting that for the SAE these variables must be available at both the small area level and within the survey – are not able to strongly predict which individuals do and do not show these outcomes. This can be explained by a combination of factors: the extent to which these health behaviours cut across these explanatory categories; a degree of pure randomness in health behaviours; and the absence of some explanatory factors of relevance to explaining health outcomes (e.g. attitudes towards health, dietary behaviours).

As noted earlier, these weak underlying predictive foundations do not *necessarily* equate to poor external validation in terms of the accuracy of the actual estimates themselves when compared with the 'true' values estimated directly from the Welsh Health Survey. However, such foundations increase the likelihood that the IPF will struggle to account for the spatial variation in the outcomes across Welsh USOAs and reduce our confidence in the likely validity[15] and stability of resulting small area estimates. For these outcomes, although all four sets of small area estimates come from underlying regression models of relatively weak power the external validation statistics show that the estimates for smoking and obesity in particular map relatively well onto the direct USOA estimates and so can be considered to have been estimated successfully and to be acceptable for policy use.

The final row of Table 15 summarises the external validation of the four sets of resulting small area estimates in the form of Pearson's correlation coefficients between the unstandardized IPF estimates at USOA scale with the USOA percentages estimated directly from weighted analyses of the Welsh Health Survey (ignoring the confidence intervals around those 'true' direct survey estimates). The IPF USOA estimates for smoking and obesity correlate highly with the point estimates derived directly from the Welsh Health Survey whilst the estimates for being physically active and, to a lesser extent, drinking above guidelines validate

[15] Where validity refers to their similarity to known 'true' (and potentially unknown) population values.

less well externally but still remain moderately strongly correlated with those direct survey estimates.

Table 15: Optimal set of constraints for each outcome variable

	Smoking	Obesity	Physically Active	Drinking above guidelines
Individual level covariates	Tenure	Limiting illness	Limiting illness	Limiting illness
	Economic Status	Economic Status	Economic status	Economic status
	Simplified NS-SEC	Simplified NS-SEC	Dependent children	Simplified NS-SEC
				Tenure
	Age-sex bands	Age-sex bands	Age-sex bands	Age-sex bands
	Highest Qualifications	Highest Qualifications	Highest Qualifications	Highest Qualifications
		Poor health	Poor health	Poor health
				Ethnic group
Area-level covariates	WIMD	WIMD	Rural/Urban	Rural/Urban
	UOAC[16]	UOAC	UOAC	Region
Pseudo-R2	10.3%	4.2%	8.0%	7.9%
External validation correlation	0.92	0.87	0.71	0.64

An alternative approach to the external validation is to ask what percentage of the 'true' direct survey central point estimates fall within the credible intervals around the central IPF point estimates (again ignoring that those 'true' direct survey values are themselves estimates with confidence intervals). Table 16 shows that in general the credible intervals produced around the IPF point estimates are relatively narrow, ranging from an average of 4.3 percentage points for the physical activity outcome up to an average of 7.6 percentage points for the drinking above guidelines outcome. The smoking and obesity results are again the most positive findings with 89% and 90% of USOA direct survey point estimates falling within their respective USOA's IPF credible intervals. The results for the drinking above guidelines estimates are also relatively strong with 83% of USOA direct survey point estimates falling within their respective USOA's IPF credible intervals. The findings for the physical activity estimates are notably weaker however with 62% of the direct survey point estimates falling within the IPF credible intervals for their USOA and 38% falling outside of those intervals, even if this outcome shows the narrowest estimated credible intervals.

[16] As with the MOAC geodemographic classification in work strand two, the UOAC classification refers to the count of the USOA adult population inside each supergroup at the LSOA level.

Table 16: Direct survey estimates in relation to IPF credible intervals

	WHS direct survey estimates within the IPF credible intervals	WHS direct survey estimates outside of the IPF credible intervals	Average width of the IPF credible intervals (percentage points)
Smokers	84	10	5.2
	89%	11%	
Obesity	85	9	7.4
	90%	10%	
Physical Activity	58	36	4.3
	62%	38%	
Drinking above guidelines	78	16	7.6
	83%	17%	

Examining the impact of smaller base survey files

The estimates discussed above in work strand five are based on a pooled Welsh Health Survey dataset covering the years 2008 to 2013 inclusive. These are the same pooled data on which the existing direct survey estimates are based. However, within this work strand there is an additional interest in exploring the potential for/impact of estimates from smaller and/or less frequent survey files. To explore these issues the IPF was replicated on the following three alternative sets of base survey files and the results compared with the IPF from the full pooled survey dataset presented above, as well as with the direct survey estimates:

1. single year data from 2008 and 2013 respectively (in order to explore the impact on estimates of using full sample sizes collected every several years);
2. a pooled dataset of total sample size roughly equivalent to one WHS survey year made up of a randomly selected one-sixth of survey cases from each Welsh Health Survey year 2008 to 2013 (in order to explore the impact on estimates of using a smaller, rolling survey approach).

Tables 17 and 18 present tables of Pearson's correlation coefficients across all of these various sets of IPF USOA estimates as well as against the direct survey estimates. As noted above, limiting illness is included amongst the optimal set of constraints for three of the outcome variables yet limiting illness is not present in the 2008 survey year. For these outcome variables, therefore, limiting illness is included as a constraint in the 2013 IPF and in the pooled six-year IPF (at the expense of approximately half the survey cases) but is not included in the 2008 IPF nor in the IPF based on a one-sixth random sample of each survey year.

Looking down the columns of Tables 17 and 18 headed 'Survey' shows the correlations between the direct survey estimates and the different sets of IPF estimates for each outcome variable. Focusing firstly on the findings for the smoking and obesity estimates in Table 17, the correlations are strong in all cases and do not vary much across the different sets of IPF estimates. For the USOA estimates of physical activity and drinking above guidelines in Table 18, however, these

correlations are both at lower levels and show greater variability across the differing base survey datasets. For the physical activity estimates the pooled survey data of all survey years and the 2008 IPF deliver correlations with the direct survey estimate of 0.71 whilst the IPF based on the 2013 WHS and on the random one-sixth sampling of all six WHS survey years deliver correlations with the direct survey estimate of 0.55. Amongst the IPF estimates of drinking above guidelines these external validation correlations vary between a low of 0.5 for the 2013 WHS IPF estimates up to 0.68 for the random one-sixth sampling of all six WHS survey years.

Table 17: Comparing the external validation of USOA smoking and obesity estimates from differently sized base surveys

% of USOA adults estimated to smoke						% of USOA adults estimated to be obese					
	Direct Survey 08 – 13	**IPF 08 – 13**	**IPF 08**	**IPF 13**	**IPF 1/6 each year**		**Direct Survey 08 – 13**	**IPF 08 – 13**	**IPF 08**	**IPF 13**	**IPF 1/6 each year**
Direct Survey 08 – 13	1.00					**Direct Survey 08 – 13**	1.00				
IPF 08 – 13	0.92	1.00				**IPF 08 – 13**	0.87	1.00			
IPF 2008	0.89	0.97	1.00			**IPF 08**	0.85	0.94	1.00		
IPF 2013	0.91	0.99	0.96	1.00		**IPF 13**	0.86	0.98	0.94	1.00	
IPF 1/6 each year	0.91	0.99	0.95	0.96	1.00	**IPF 1/6 each year**	0.85	0.95	0.97	0.94	1.00

Table 18: Comparing the external validation of USOA estimates of physical activity and drinking above guidelines from differently sized base surveys

% of USOA adults estimated to do adequate physical activity						% of USOA adults estimated to drink above guidelines					
	Direct Survey 08 – 13	**IPF 08 – 13**	**IPF 08**	**IPF 13**	**IPF 1/6 each year**		**Direct Survey 08 – 13**	**IPF 08 – 13**	**IPF 08**	**IPF 13**	**IPF 1/6 each year**
Direct Survey 08 – 13	1.00					**Direct Survey 08 – 13**	1.00				
IPF 08 – 13	0.71	1.00				**IPF 08 – 13**	0.64	1.00			
IPF 2008	0.71	0.80	1.00			**IPF 08**	0.57	0.60	1.00		
IPF 2013	0.55	0.78	0.67	1.00		**IPF 13**	0.50	0.86	0.42	1.00	
IPF 1/6 each year	0.55	0.90	0.58	0.72	1.00	**IPF 1/6 each year**	0.68	0.84	0.78	0.66	1.00

Evaluating the healthy lifestyles estimates

Across these various results there are both real positive findings as well as notes of caution required. In terms of the positives, the estimates for smoking and obesity in particular validate well against the direct survey estimates, despite being based on relatively weak underlying predictive models. This is interesting in that it highlights the *potential* for small area estimation to produce acceptably accurate small area estimates even in these circumstances, even if one would wish to ensure that this

was the case in such cases via a robust process of external validation. A second notable finding is the ability of the small area estimation to produce similar estimates with dramatically smaller base survey sizes, with implications in terms of increasing the efficiency (e.g. greater ability to rotate the outcome variables of interest collected each survey year) and reducing the cost of future survey data collection.

Amongst these promising findings there is also need for caution. Looking across the full set of results taken together, and seeing them as a limited form of repeat testing, raises questions around the confidence to be placed in some of these small area estimates. In particular, differences in performance are evident across the four sets of estimated outcomes (Tables 15 and 16) as well as across different survey base files for the outcomes of physical activity and drinking above guidelines (Table 18).

Various potential explanations for these differences exist – differential model power, the spatial variation in the outcomes, small changes in the levels of the outcomes over time, the constraints used – but these possibilities seem unable to account for the patterns seen. The underlying model power does not seem decisive to either the level or consistency of the external validation statistics and all four outcomes show relatively similar extents of spatial variation across the USOAs. The four outcomes do show varying levels across Welsh adults (Figure 15) that do map onto the quality of the external validation statistics but this is not seen as an adequate explanation for this variation in results.

Rather, although the validation of the smoking and obesity estimates in particular appears satisfactory, consideration of the results as a whole across this work strand suggests caution around the reliance on small area estimation techniques based on underlying models of weak predictive power unless one is able to see sufficiently powerful external validation statistics of the resulting estimates against existing known data, preferably at the target small area scale. The result of basing the small area estimation on relatively weak predictive foundations is that even though the IPF effectively reweights survey cases to match the USOA profiles across the constraints the links between those constraints (and, hence, the reweighted survey cases) and the outcome variables is weak, with relatively little able to be explained stably and systematically from the constraint variables. The implication is that resulting estimates are susceptible to volatility dependent upon the types of outcomes that the reweighted survey cases in that survey dataset happen to take. With stronger predictive relationships between characteristics and outcomes the small area estimation would be better able to offer greater systematic explanatory power to outcomes and hence would be more likely to produce estimates that are closer to the ‘true’ (but typically unknown) population values as well as that are more stable across different base survey datasets. At the same time, weaker predictive foundations *can* deliver small area estimates that are close to ‘true’ values as is seen with the smoking and obesity estimates in particular.

Therefore, although these results demonstrate that one can *potentially* create small area estimates of these healthy lifestyle outcomes that validate well externally, the variability in results both between and within these four sets of external validation statistics suggests that the quality of the resulting small area estimates cannot

necessarily be relied upon nor predicted in advance. On this basis we would urge caution around reliance upon small area estimation as a means to create spatially detailed information where outcomes are able to be only weakly predicted in underlying regression models. In raising the possibility that acceptable small area estimates can *potentially* be delivered even from these weak predictive foundations, however, these analyses also raise the possibility for future research to seek to identify the conditions under which this conversion between underlying model power and the quality of the external validation is maximised in order that clearer guides can be established around the likely effectiveness of small area estimation from underlying models of varying predictive power.

Profiling the areas where IPF estimates fit less well

As with any estimation process, the IPF estimates for some areas fit less well with known data. As part of the project, an investigation was undertaken of the characteristics of areas for which the estimates fit less well.

There are three separate strands of estimates:

- Personal internet use and household internet access estimates: these are MSOA estimates based on individual factors only and are aggregated up to LA level and compared with the central direct survey point estimates
- The optimal/selected Welsh speaking and poor health estimates compared with known Census 2011 values at MSOA level
- The four healthy lifestyles estimates from the WHS that are compared at USOA level with direct survey estimates (based on the random 1/6 draw of the 2008-2013 WHS)

The main finding was that it is the level of the outcome variable that affects whether the estimates fit poorly. Where areas have particularly low or high values of the outcome, based on the 'known' data (e.g. Census or direct survey estimates), the small area estimates are less likely to fit well. Where the known outcome value is particularly high, the small area estimate tends to be lower than that value; and where the known outcome value is low, the small area estimate tends to be higher than that value. This can help to identify areas where poorer estimation would be expected (high/low areas on the outcome) as well as the characteristics of those areas (factors associated with those high/low values, which varies depending on the outcome).

It is doubtful that any method would completely adjust the lack of fit for lower/higher outcome variables as this is to some extent a natural consequence of any modelled estimation process. However, selecting only cases from same region (as discussed elsewhere in this report) does help. The post-estimation constraining of estimates to known higher geographical levels will also help to pull these tails up/down as well as fit to accepted higher scale totals, especially ones that take into account the estimated level of the outcome. This lends weight to choosing the implementation of the constraining process that takes into account the level of the outcome.

Conclusions and Recommendations

The five work strands of this project have focused on methodological experimentation and learning around how best to maximise the performance of potential future Welsh Government small area estimation work using the IPF approach. This experimentation has generated a series of new small area estimates at a variety of spatial scales around internet use and access, poor health, Welsh speaking and healthy lifestyles. It has also offered a methodological template for the small area estimation of these and other outcomes in future. Overall the project demonstrates that IPF is a viable methodological approach for the production of accurate small area estimates and accompanying credible intervals where outcomes are able to be modelled with a reasonable degree of predictive power. The small area estimates generally validate well against known external data, including at the demanding (and unusually small) target small area scale for the Welsh speaking and poor health outcomes where data are known at small area level from Census 2011. The testing of the credible intervals shows that they function well and give a reasonably solid indication of the likely range of the 'true' underlying population values.

Careful thought throughout needs to be placed on the specification of the small area estimation approach in order to maximise the quality of the resulting small area estimates, including consideration of the optimal combination of both individual and area-level constraints. These will vary depending upon the source of the variation in the outcomes being estimated – spatial variation matters for the estimation of Welsh speaking, for example, but not for internet use or poor health – and will affect both the estimation of the point estimates and credible intervals. The inclusion of dummy variables relating to the categories on these area-level variables within the initial survey regression modelling can help to identify their relative importance to the separate outcomes. The consideration of the optimal set of constraints should include not only consideration of their predictive power but also sensitivity to potential over fitting through the incorporation of too many constraints and/or constraints that are too sparsely populated so that the IPF struggles to populate every cross-classified combination of the constraint factors.

Post-estimation constraining to known 'true' values at higher scales is rarely used within small estimation work at present but, despite the assumptions inherent within the approach, is shown to offer a potential additional step to enhance the quality of the small area estimates produced.

The attempts to produce estimates relating to healthy lifestyles highlights the potential for small area estimation to produce acceptable estimates even in situations where the outcome is weakly predicted, although robust external validation is critical in such circumstances given the greater potential for poorer estimates. This work also highlights the ability to produce comparable small area estimates from dramatically smaller base survey files, opening up possibilities around increased efficiencies (e.g. ability to focus on different outcome variables each year rather than repeat collection of the same outcomes) and reduced costs in future survey data collection. At the same time, it is important to recognise that the limited ability of the

underlying predictive models to explain these outcomes necessarily introduces a need for greater caution in approaching the estimation of these outcomes and this is evident in this case in the variations in the external validation statistics seen across these outcomes and across their alternative base survey specifications.

The successful validation of estimates across the project gives confidence in the ability to effectively conduct small area estimation of these and additional outcomes in future, particularly in cases where those outcomes can be explained with a reasonable degree of predictive power – around 30% as a suggested minimum benchmark for the predictive power of the underlying regression models. Moreover, it is possible to expand the types of outcomes estimated down to small area level if specific insights are desired. For example, whilst this project focuses in parts on the estimation down to small area level of the percentage of adults that self-report as speaking Welsh, the National Survey contains further variables relating to specific aspects of Welsh speaking: frequency of speaking Welsh; Welsh language ability; and different configurations of skills in reading, writing and/or speaking the Welsh language. Given that these variables are captured in the National Survey along with explanatory variables able to explain an adequate share of their overall variance these more specific dimensions of Welsh language ability could become target outcome variables for future work if desired, even if direct external validation at the target small area level would not be possible given their absence from the Census data. The same is true of other survey variables of potential interest, assuming similar ability to predict given the explanatory factors also captured in national surveys, and offers rich terrain for further spatially detailed insights from small area estimation across a range of potential outcomes of interest.

The project findings raise several **recommendations** for ways to further develop the performance of small area estimation work:

- As is typically the case with small area estimation work, this project relies upon Census data to describe the characteristics of the small areas to which the IPF reweights the survey cases. Given that Census data are collected only every decade, and indeed that there are questions over the continued existence of the UK Census in the long term, it would be valuable to explore the viability of bringing together data from non-Census sources (e.g. administrative or commercial data) to give small area covariate data that are timely, comprehensive and sufficiently rich;

- In reweighting the survey cases to match the small area profiles the IPF places no restrictions on either the extent to which weights can change nor on the maximum size that weights can take. Whilst this provides the IPF with greater flexibility for the reweighting it may also expose that reweighting to the reliance of large weights on some survey cases. Other spatial microsimulation approaches such as GREGWT do incorporate the ability to impose such conditions on the reweighting, but this is not something that has ever been explored in the IPF approach;

- The project makes use of an innovative approach for the creation of credible intervals around the central point estimates from the IPF. In doing so the project responds to the key weakness of spatial microsimulation approaches to small area estimation which is their current inability to provide such estimates of variance. Future work could valuably examine the statistical foundations of the suggested approach to building the credible intervals and this may broaden out to a full statistical understanding of the IPF approach itself;

- Further research exploring the conditions under which the conversion between underlying model power and the quality of the external validation is maximised would be beneficial. This would allow clearer guides to be established for the likely effectiveness of small area estimation from underlying models of varying predictive power.

References

Anderson, B. (2007) Creating *small area income estimates for England: spatial microsimulation modelling*, a report to the Department of Communities and Local Government. London: Department of Communities and Local Government.

Anderson, B. (2012) 'Estimating Small Area Income Deprivation: An Iterative Proportional Fitting Approach' in Edwards, K. and Tanton, R. (eds) *Spatial Microsimulation: A Reference Guide for Users*. London: Springer.

Bajekal, M., Scholes, S., Pickering, K. and Purdon, S. (2004) *Synthetic estimation of healthy lifestyle indicators: Stage one report*. London: National Centre for Social Research.

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B. and Rossiter, D. (2005) 'SimBritain: A Spatial Microsimulation Approach to Population Dynamics', *Population, Space and Place*, 11, 13-34.

Birkin, M. and Clarke, G. (2012) 'The enhancement of spatial microsimulation models using geodemographics', *Annals of Regional Science*, 49, 515-532.

Heady, P., Clarke, P., Brown, G., Ellis, K., Heasman, D., Hennell, S., Longhurst, J. and Mitchell, B. (2003) *Model based area estimation series No. 2: Small area estimation project report*. London: Office for National Statistics.

Ipsos MORI (2015) *Multi-level modelling and small area estimation work for the National Survey for Wales.* London: Ipsos MORI

Marshall, A. (2010) *Small area estimation using ESDS government surveys – An introductory guide*. Economic and Social Data Service.

Rahman, A. (2008) *A review of small area estimation problems and methodological developments.* University of Canberra: NATSEM Discussion Paper Issue 66.

Smith, D., Clarke, G. and Harland, K. (2009) 'Improving the synthetic data generation process in spatial microsimulation models', *Environment and Planning A,* 41, 1251-1268.

Tanton, R., Williamson, P. and Harding, A. (2014) *Comparing two methods of reweighting a survey file to small area data'*, International Journal of Microsimulation, 7(1),pp76-99.

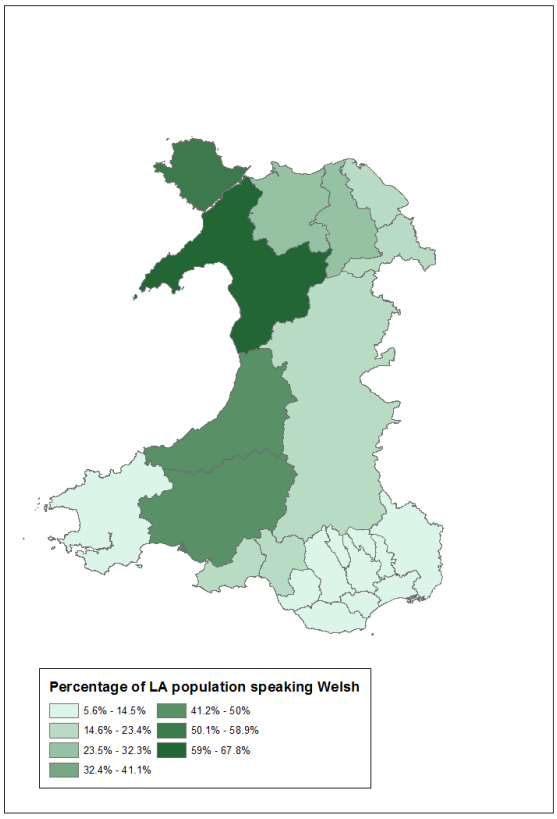
Whitworth, A. (2013) (ed.) *Evaluations and improvements in small area estimation methodologies.* Economic and Social Research Council: National Centre for Research Methods methodological review paper.

Whitworth, A., Jones, P. and Thomas, B. (2015) *Understanding Wales at the small area level: Small area estimation*. Cardiff: Welsh Government.

Williamson, P. (2013) ‘An evaluation of two synthetic small-area microdata simulation methodologies: synthetic reconstruction and Combinatorial Optimisation’ in Tanton, R. and Edwards, K. (eds) *Spatial microsimulation: a reference guide for users*. London: Springer.

Appendix

Figure 14: Bespoke design of regions for the tailoring of survey case selection for the estimation of Welsh speaking



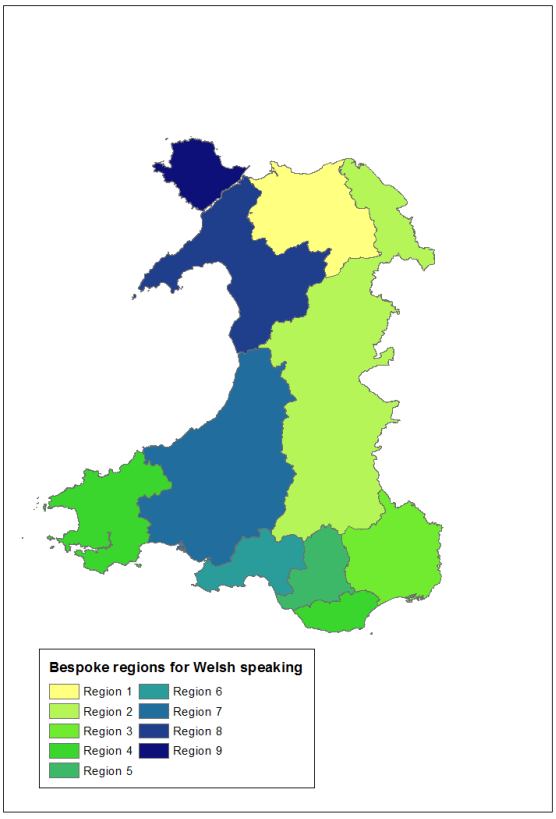
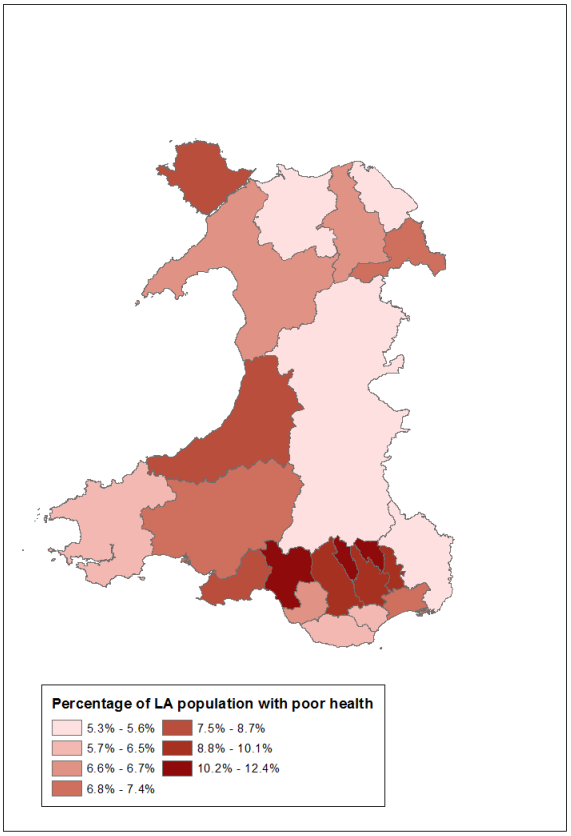


Figure 15: Bespoke design of regions for the tailoring of survey case selection for the estimation of health outcomes



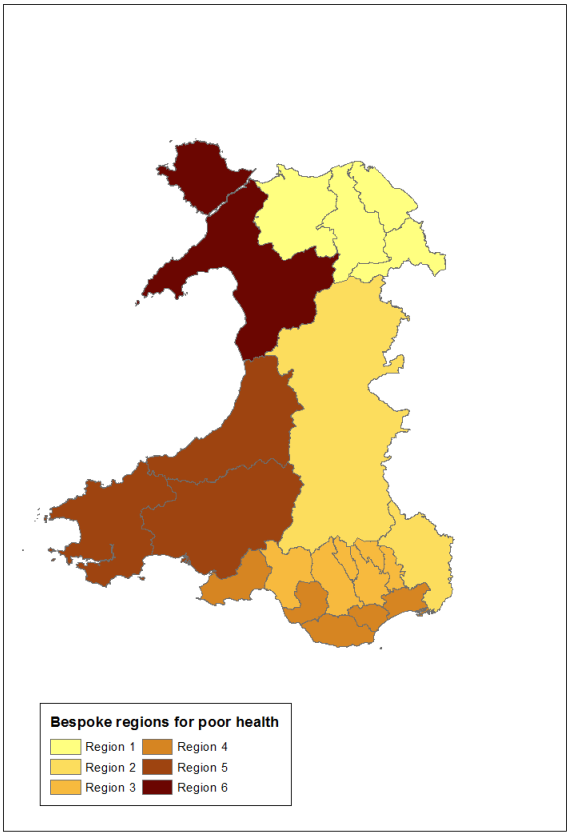


Table 19: Internal validation statistics for the percentage of adults in poor health

(unshaded cells related to constraints used in the IPF; shaded cells relate to variables not included as constraints in the IPF)

	Mean standardised error								% of small area estimates more than 20% from Census counts						
	Strand 1: Individual only constraint		Strand 2: Individual plus area constraints						Strand 1: Individual only constraint		Strand 2: Individual plus area constraints				
	LSOA	MSOA	Plus MOAC	Plus rural /urban	Plus region	Plus WIMD	Plus WIMD &		LSOA	MSOA	Plus MOAC	Plus rural /urban	Plus region	Plus WIMD	Plus WIMD &
F1629	0.0	0.0	0.2	0.0	0.0	0.0	0.0		0.0	0.0	0.2	0.0	0.0	0.0	0.0
F3049	0.0	0.0	0.0	0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0	0.0	0.0	0.0
F5064	0.0	0.0	0.3	0.0	0.0	0.0	0.0		0.0	0.0	0.2	0.0	0.0	0.0	0.0
F65up	0.0	0.0	0.3	0.0	0.0	0.0	0.0		0.0	0.0	0.2	0.0	0.0	0.0	0.0
M1629	0.0	0.0	0.3	0.0	0.0	0.0	0.0		0.0	0.0	0.2	0.0	0.0	0.0	0.0
M3049	0.0	0.0	0.1	0.0	0.0	0.0	0.0		0.0	0.0	0.2	0.0	0.0	0.0	0.0
M5064	0.0	0.0	0.3	0.0	0.0	0.0	0.0		0.0	0.0	0.2	0.0	0.0	0.0	0.0
M65up	0.0	0.0	0.0	0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0	0.0	0.0	0.0
No limiting illness	0.0	0.0	0.0	0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0	0.0	0.0	0.0
Limiting illness	0.0	0.0	0.2	0.0	0.0	0.0	0.0		0.0	0.0	0.2	0.0	0.0	0.0	0.0
In work	0.2	0.2	0.3	0.2	0.2	0.2	0.2		0.0	0.0	0.0	0.0	0.0	0.0	0.0
Inactive	0.2	0.2	0.3	0.2	0.2	0.2	0.2		0.0	0.0	0.0	0.0	0.0	0.0	0.0
Unemployed	0.2	0.3	0.6	0.2	0.2	0.3	0.3		0.0	0.0	0.2	0.0	0.0	0.0	0.0
Retired	0.6	0.7	1.2	0.6	0.6	0.6	0.6		0.0	0.0	0.2	0.0	0.0	0.0	0.0
Student	0.3	0.3	0.4	0.3	0.3	0.3	0.3		0.0	0.0	0.0	0.0	0.0	0.0	0.0
No quals	0.3	0.3	0.6	0.3	0.3	0.3	0.3		0.0	0.0	0.2	0.0	0.0	0.0	0.0
L1 quals	0.1	0.1	0.1	0.1	0.1	0.1	0.1		0.0	0.0	0.0	0.0	0.0	0.0	0.0
L2 quals	0.1	0.1	0.1	0.1	0.1	0.1	0.1		0.0	0.0	0.0	0.0	0.0	0.0	0.0
L3 quals	0.2	0.2	0.5	0.2	0.2	0.2	0.2		0.0	0.0	0.2	0.0	0.0	0.0	0.0
L4+ quals	0.1	0.1	0.1	0.1	0.1	0.1	0.1		0.0	0.0	0.0	0.0	0.0	0.0	0.0
Owner occupier	0.0	0.0	0.0	0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0	0.0	0.0	0.0
Private rent	0.0	0.0	0.0	0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0	0.0	0.0	0.0
Social rent	0.0	0.0	0.3	0.0	0.0	0.0	0.0		0.1	0.0	0.2	0.0	0.0	0.0	0.0
NS-SEC Higher	12.0	10.2	9.9	10.7	10.5	9.6	10.1		16.7	12.9	8.0	13.2	10.0	8.3	8.8
NS-SEC Intermediate	13.5	12.1	13.6	12.6	13.6	10.5	12.0		21.5	18.3	15.4	13.9	18.3	13.2	19.3
NS-SEC manual	16.9	14.2	10.8	13.6	13.1	9.7	9.7		23.9	20.0	13.7	17.8	18.0	12.7	12.0
Car	7.7	6.9	6.8	6.6	6.6	6.2	6.1		1.9	0.5	2.9	1.5	1.2	0.0	1.2
Dependent kids	18.9	11.1	7.0	11.2	10.6	10.8	9.2		14.5	6.1	5.6	6.1	6.3	5.9	6.8

Table 20: Internal validation statistics for the percentage of adults speaking Welsh

(unshaded cells related to constraints used in the IPF; shaded cells relate to variables not included as constraints in the IPF)

	Mean standardised error								% of small area estimates more than 20% from Census counts						
	Strand 1: Individual only constraint		Strand 2: Individual plus area constraints						Strand 1: Individual only constraint		Strand 2: Individual plus area constraints				
	LSOA	MSOA	Plus MOAC	Plus rural	Plus region	Plus WIMD	Plus WIMD & region		LSOA	MSOA	Plus MOAC	Plus rural	Plus region	Plus WIMD	Plus WIMD & region
F1629	0.0	0.0	0.2	0.0	0.0	0.0	1.4		0.0	0.0	0.2	0.0	0.0	0.0	1.7
F3049	0.0	0.0	0.3	0.0	0.0	0.0	1.9		0.0	0.0	0.2	0.0	0.0	0.0	2.4
F5064	0.0	0.0	0.3	0.0	0.0	0.0	1.5		0.0	0.0	0.2	0.0	0.0	0.0	1.0
F65up	0.0	0.0	0.3	0.0	0.0	0.0	1.5		0.0	0.0	0.2	0.0	0.0	0.0	1.7
M1629	0.0	0.0	0.3	0.0	0.0	0.0	2.0		0.0	0.0	0.2	0.0	0.0	0.0	2.9
M3049	0.0	0.0	0.1	0.0	0.0	0.0	1.1		0.0	0.0	0.2	0.0	0.0	0.0	1.7
M5064	0.0	0.0	0.3	0.0	0.0	0.0	1.5		0.0	0.0	0.2	0.0	0.0	0.0	1.0
M65up	0.0	0.0	1.0	0.0	0.0	0.0	2.2		0.0	0.0	0.2	0.0	0.0	0.0	2.0
Id: Other	0.0	0.0	0.0	0.0	0.0	0.0	1.0		0.0	0.0	0.0	0.0	0.0	0.0	1.0
Id: Rest UK	0.0	0.0	0.0	0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0	0.0	0.0	0.0
Id: Welsh-Brit	0.0	0.0	0.0	0.0	0.0	0.0	0.2		0.0	0.0	0.0	0.0	0.0	0.0	0.2
Id: Welsh	0.0	0.0	0.0	0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0	0.0	0.0	0.0
Born Ireland	0.1	0.0	0.6	2.3	0.1	0.1	9.7		3.7	0.0	1.2	2.2	0.0	0.0	9.3
Born England	0.0	0.0	0.4	0.0	0.0	0.0	1.4		0.0	0.0	0.2	0.0	0.0	0.0	1.2
Born N.Irl	0.0	0.0	8.3	0.0	0.1	0.0	29.1		6.3	0.0	8.3	0.0	0.0	0.0	29.0
Born Scotland	0.0	0.0	3.7	0.0	0.0	0.0	8.5		0.3	0.0	3.7	0.0	0.0	0.0	8.0
Born Wales	0.0	0.0	0.1	0.0	0.0	0.0	0.5		0.0	0.0	0.0	0.0	0.0	0.0	0.0
Not UK born	0.0	0.0	0.4	0.1	0.1	0.0	2.7		0.0	0.0	0.2	0.0	0.0	0.0	3.4
No quals	0.0	0.0	0.3	0.0	0.0	0.0	2.2		0.0	0.0	0.5	0.0	0.0	0.0	3.7
L1 quals	0.0	0.0	0.2	0.0	0.0	0.0	2.8		0.0	0.0	0.2	0.0	0.0	0.0	3.7
L2 quals	0.0	0.0	0.1	0.0	0.0	0.0	1.7		0.0	0.0	0.0	0.0	0.0	0.0	2.9
L3 quals	0.0	0.0	0.3	0.0	0.0	0.0	1.7		0.0	0.0	0.2	0.0	0.0	0.0	2.7
L4+ quals	0.0	0.0	0.3	0.0	0.0	0.0	1.9		0.0	0.0	0.2	0.0	0.0	0.0	3.2
Car access	0.0	0.0	0.0	0.0	0.0	0.0	0.2		0.0	0.0	0.0	0.0	0.0	0.0	0.2
No car access	0.0	0.0	0.1	0.0	0.0	0.0	0.4		0.0	0.0	0.0	0.0	0.0	0.0	0.2
NS-SEC higher	0.0	0.0	0.1	0.0	0.0	0.0	0.7		0.0	0.0	0.2	0.0	0.0	0.0	1.0
NS-SEC Intermediate	0.0	0.0	0.0	0.0	0.0	0.0	0.9		0.0	0.0	0.0	0.0	0.0	0.0	1.5
NS-SEC manual	0.0	0.0	0.1	0.0	0.0	0.0	1.0		0.0	0.0	0.0	0.0	0.0	0.0	2.2
NS-SEC never worked	0.0	0.0	0.4	0.0	0.0	0.0	2.9		0.0	0.0	0.2	0.0	0.0	0.0	4.9
No limiting illness	4.0	3.2	3.1	3.1	3.2	2.8	4.2		0.1	0.0	0.2	0.0	0.0	0.0	0.5
Has limiting illness	11.3	8.6	8.9	8.7	8.9	7.8	13.3		12.6	3.9	8.0	7.1	7.3	2.7	14.6
In work	8.3	6.4	6.4	7.2	6.3	6.7	6.9		3.5	4.1	3.9	3.9	4.1	4.1	4.4
Inactive	44.5	41.6	36.7	40.4	40.3	34.8	37.1		63.8	68.5	69.0	67.6	71.2	71.0	69.8
Retired	4.7	5.6	6.7	5.5	5.6	5.3	7.8		0.8	0.0	3.2	0.0	0.0	0.0	5.4
Student	28.0	26.8	29.2	28.5	29.5	25.3	37.9		62.5	62.4	71.0	60.7	65.4	62.0	72.9
Unemployed	25.3	20.2	22.8	20.5	25.0	17.8	36.1		40.6	36.8	43.9	40.5	40.2	35.9	61.0

Table 21: Internal validation statistics for the percentage of adults using the internet and with internet access

(unshaded cells related to constraints used in the IPF; shaded cells relate to variables not included as constraints in the IPF)

	Mean standardised error				% of small area estimates more than 20% from Census counts			
	Internet Use		Internet Access		Internet Use		Internet Access	
	LSOA	MSOA	LSOA	MSOA	LSOA	MSOA	LSOA	MSOA
Car access	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
No car access	0.2	0.3	0.1	0.1	0.0	0.0	0.0	0.0
F1629	0.0	0.0	0.3	0.4	0.0	0.0	0.0	0.0
F3049	0.0	0.0	0.4	0.5	0.0	0.0	0.0	0.0
F5064	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0
F65up	0.0	0.0	0.9	1.2	0.0	0.0	0.0	0.0
M1629	0.0	0.0	0.3	0.3	0.0	0.0	0.0	0.0
M3049	0.0	0.0	0.3	0.4	0.0	0.0	0.0	0.0
M5064	0.0	0.0	0.2	0.3	0.0	0.0	0.0	0.0
M65up	0.0	0.0	0.8	1.1	0.0	0.0	0.0	0.0
In work	0.6	0.8	0.8	0.4	0.0	0.0	0.0	0.0
Inactive	0.6	0.9	0.7	0.3	0.0	0.0	0.0	0.0
Retired	2.1	2.8	2.2	1.0	0.0	0.0	0.0	0.0
Student	0.7	1.1	1.1	0.5	0.0	0.0	0.0	0.0
Unemployed	0.7	1.0	0.8	0.4	0.0	0.0	0.0	0.0
No quals	0.1	0.1	0.3	0.4	0.0	0.0	0.0	0.0
L1 quals	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0
L2 quals	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0
L3 quals	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0
L4+ quals	0.1	0.1	0.2	0.2	0.0	0.0	0.0	0.0
Lone parent	0.7	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Married Pensioner	2.2	3.2	0.0	0.0	0.0	0.0	0.0	0.0
Single Pensioner	2.3	3.2	0.0	0.0	0.0	0.0	0.0	0.0
Single Working Age	0.6	0.9	0.0	0.0	0.0	0.0	0.0	0.0
Two adults no kids	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Two adults w/ kids	0.7	1.0	0.0	0.0	0.0	0.0	0.0	0.0
NS-SEC higher	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0
NS-SEC intermed	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NS-SEC routine	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
NS-SEC never worked	0.1	0.2	0.1	0.1	0.0	0.0	0.0	0.0
Owner occupier	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0
Private rent	0.6	0.9	0.2	0.3	0.0	0.0	0.0	0.0
Social rent	0.2	0.2	0.1	0.1	0.1	0.0	0.1	0.0
Has limiting illness	10.4	8.7	10.4	8.7	10.7	4.6	10.8	4.6
No limiting illness	3.9	3.4	3.9	3.4	0.1	0.0	0.1	0.0
Dependent kids	13.7	8.0	13.1	8.7	4.7	2.9	4.6	2.9
No dependent kids	3.8	2.9	3.5	3.3	1.8	1.0	1.7	1.0
White British	4.0	3.7	4.0	3.7	4.0	3.4	4.0	3.4

Technical Appendix: Example segment of Stata syntax to conduct IPF

```
/* Open small area covariate file showing one row per target small area (eg MSOAs) with aggregate totals on each constraint
group as variables and just keep the one MSOA case to run this run of IPF on (MSOA1 in this example)*/
use     "$msoa/msoa covariate base file_master.dta", clear
keep if _n ==1
sort key
saveold "$msoa/msoa covariate 1.dta", replace

/* Open the survey base file and merge in that one target MSOA area */
use   "$raw/survey base file.dta", clear
merge m:1 using "$msoa/msoa covariate 1.dta"
tab   _m
drop  _m

/* Set starting weight value */
generate wt1 = WalesAdultWeight

/****** Reweight on constraint 1: Car Access *****/

/* Generate weighted survey totals based on current weight */
generate yescar_wt=yescar * wt1
generate nocar_wt=nocar * wt1
egen yescar_s=total(yescar_wt)
egen nocar_s=total(nocar_wt)
drop *_wt

/* Reweight survey cases based on census total/survey total on constraint 1 (car access) */
generate wt2 = .
replace wt2=wt1 * (yescar_c / yescar_s if yescar==1
replace wt2=wt1 * (nocar_c / nocar_s if nocar==1

/****** Reweight on constraint 2: Dependent Children *****/

/* Generate new weighted survey totals based on the updated current weight */
generate depkid_wt=depkid * wt2
generate nodepkid=nodepkid * wt2
egen depkid_s=total(depkid_wt)
egen nodepkid_s=total(nodepkid_wt)
drop *_wt

/* Reweight survey cases based on census total/survey total on constraint 2 (dep kids) */
generate wt3 = .
replace wt3=wt2 * (depkid_c / depkid_s if depkid==1
replace wt3=wt2 * (nodepkid_c / nodepkid_s if nodepkid==1

…….. continue through rest of constraints (and then loop back over them) until complete……

/* Rename the last weight as the final weight & only keep relevant variables */
rename w81 finalwt
keep finalwt lsoacode internetuse

/* Use final weight to calculate a weighted average as MSOA's estimate of internet use */
generate internetuse_wt = internetuse * finalwt
collapse (mean) finalwt internetuse_wt, by(msoacode)
save "$work/internetuse msoacode1.dta",replace

/* Move onto the next target small area (e.g. MSOA2) */
```

March 2015

Multi-level Modelling and Small Area Estimation Work for the National Survey for Wales

Prepared for the Welsh Government

Sarah Tipping and John D'Souza

Contents

1 Background

The Welsh Government is investigating methods for generating small area estimates for the National Survey for Wales and the Welsh Health Survey (WHS). Ipsos MORI was commissioned to carry out a short piece of work to investigate the use of multi-level modelling techniques as an alternative to Iterative proportional fitting (IPF). The results of this work will be used alongside the results from the main C2/2014-15 contract and estimates produced by the two methods compared to provide advice on the suitability of each method.

1.1 Aims and objectives

The aims of this piece of work were:

- Run multilevel modelling in order to gain a better understanding of the variance structure of the National Survey data.
- Produce small area estimates for the 410 Welsh Medium Super Output Areas (MSOA). The estimates will be generated by fitting a multi-level logistic regression model using the 2013-14 National Survey results.
- Validate the estimates as appropriate, including by looking at intra-class correlations.

Estimates were produced for two topics: general health status and the proportion of Welsh speakers. In order to allow direct comparisons it was essential that the outcome variables were coded in the same way as those used in the main small area estimation contract. The two binary outcome variables were therefore:

- Welsh Speaker coded as 1=can speak at least a little Welsh, 0=other
- Poor health coded as 1=poor, very poor health, 0=other

Missing values were excluded. The Stata code used to derive the variables is given in Appendix A.

Whilst the outcome variables were fixed, the selection of covariates was independent. This is discussed further below.

2 Description of method

This section outlines the process used to generate the small area estimates.

2.1 Setting up covariates

There first step was to derive and select the covariates. Unlike IPF, the data used to generate the model need to be in the same format as the population data. This means we are unable to use individual-level covariates (as was used for the IPF) as the population data would not be available at individual-level (for example, it would require us to have access to individual-level Census data). Therefore we do not use variables such as whether or not the respondent has access to a car, but instead use area-level variables such as the proportion of households in the MSOA having access to a car.

The exception to this is age and sex, since many Census data releases are broken down by age and sex, the population information would be available.

Potential covariates were generated using Census data, ACORN, the Welsh Index of Multiple Deprivation and urban/rural classification aggregated to MSOA level. A full list of the variables considered is given in Appendix B.

There were a large number of potential covariates, many of which were highly correlated. Factor Analysis and correlations were used to reduce these to a more manageable number (see Appendix B).

2.2 Modelling the data

The MSOA-level variables were merged to the National Survey and were used in a multi-level model (xtlogit in Stata v13) to predict the two outcomes; Welsh speaking and poor health. This command computes random intercepts models. The MSOA is specified as the first level, i.e. the level in which respondents are clustered, and the model calculates the variance accordingly.

Only significant variables were retained in the model; variables that did not have a significant relationship with the outcome were dropped. The best set of predictors was retained for each outcome, with the result that a different set of covariates was used in each model. The covariates in the Welsh speakers model were:

- % of households in the MSOA without a car,
- % of usual residents in the MSOA who were born in Wales,
- % of usual residents in the MSOA who were married,
- % of households in the MSOA that contained a single resident,
- % usual residents in the MSOA whose highest qualification is level 4 (i.e. above A level), and
- an interaction between Assembly Economic Fora Area and % born in Wales.
- Local Authority (22 categories),
-

The covariates used in the Poor Health model were:

- An identifier for Communities First areas,
- % households in MSOA who are Group 1 (Affluent Achievers) in the 6 category ACORN code,
- % households containing dependent children, and
- % usual residents with limiting long term illness.

The full model output is given in Appendix C.

2.3 Generating small area estimates

The coefficients from each model were saved and used to create simulated values of the parameter estimates and error term. These simulated values were random realizations of the model parameter estimates, generated using the mean and covariance matrix from the model. The simulations were run 50,000 times for each outcome, giving 50,000 simulated coefficients per outcome.

These simulated parameter estimates were matched back into the population data (a file containing all 410 MSOAs with Census data, ACORN, IMD and urban indicators attached) and used to generate the small area estimates. The steps were as follows:

- Take the simulated values of the model parameters; the constant (const-i), coefficients (beta_1-i, beta_2_i, ..., etc) and error term (sigma_u-i) for i = 1 ... 50,000.

- In a loop merge in each MSOA, one at a time, and generate 50,000 estimates of xb-i = const-i + beta_1-i*x_1 + ... + Z * sigma_u-i where Z is N(0,1)

- Convert the xb-i to predicted probabilities (p-i) by taking an inverse logit. This gives us 50,000 predicted probabilities for each MSOA. The mean of these predicted probabilities is the small area estimate. In addition to the mean, the standard deviation, 2.5th percentile and 97.5th percentile for each MSOA are saved.

So, for the Welsh speaking outcome, the last few variables in the file look like this:

	SArea_20	SArea_21	SArea_22	Snocar	SbornW~s	Smarried	Shh1per	Squal4	Sconst	Slnsig2u	Ssigu
1.	-4.080	-3.305	-3.511	-0.077	0.032	-0.046	0.038	0.008	0.701	-1.489	0.475
2.	-3.280	-2.634	-3.046	-0.082	0.034	-0.047	0.052	0.001	-0.065	-1.430	0.489
3.	-3.489	-3.061	-3.589	-0.062	0.036	-0.047	0.032	0.012	0.076	-1.954	0.376
4.	-3.486	-3.096	-2.696	-0.047	0.027	-0.042	0.027	0.022	-0.200	-1.575	0.455

For the first MSOA in Local Authority 22, the fitted part on the first simulation would be:

xb1 = -3.511-0.077*nocar+ 0.032*bornWales -0.046*married + 0.038*hh1per + 0.008*qual4 + 0.701

and generate p1 as:

p1 = invlogit(xb1 + Z*0.475)

Where Z is a normal random variable. We create p1 – p50000. The mean of these 50,000 simulations is the small area estimate. The standard deviation is also saved, as well as the 2.5th and 97.5th percentiles, which are used as the lower and upper bounds.

2.4 Validating the estimates

The estimates were compared against the Census 2011 data on Welsh speaking and poor health (note, for obvious reasons, these variables were not included in the modelling).

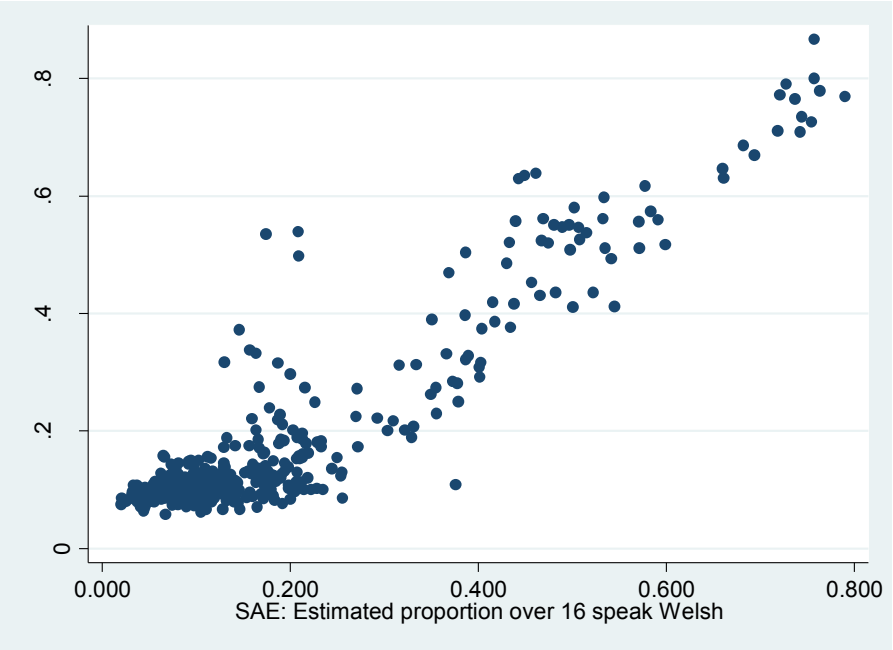
Welsh speaking

The small area estimates of Welsh speaking and the Census 2011 had a Pearson's correlation coefficient of 0.922. The means were close, although the Census proportions had a slightly larger minimum and maximum. However, note that the census variable measures the proportion of people over three who speak Welsh, rather than over sixteen, so this is only an indication.

Measure	Mean	Standard Deviation	Minimum	Maximum
Census 2011: % people aged 3yrs+ in MSOA who can speak some Welsh	0.1908	0.17012	0.0582	0.8660
Small area estimate: % adults aged 16yrs+ in MSOA who can speak at least a little Welsh	0.1914	0.16732	0.0199	0.7904

The distribution of the two estimates was checked using a scatter plot. This is shown in Figure 1. There are a small number of outliers but the two are generally close.

Figure 1: Scatter plot of small area estimate of Welsh speaking and Census data



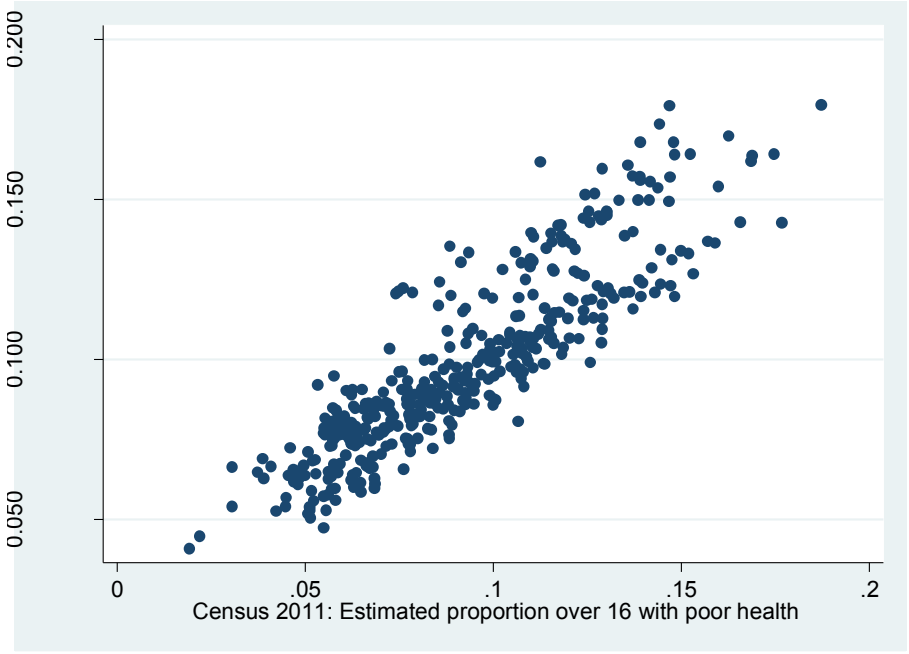
Poor Health

The small area estimates for poor health were compared to data from Census 2011. The two sets of estimates had a Pearson's correlation coefficient of 0.885, slightly lower than that for Welsh speaking. The small area estimates have a narrower range.

Measure	Mean	Standard Deviation	Minimum	Maximum
Census 2011: % people aged 16yrs+ in MSOA who have 'bad' or 'very bad' general health	0.0921	0.03085	0.0192	0.1872
Small area estimate: % adults aged 16yrs+ in MSOA who have 'poor' or 'very poor' health	0.0985	0.02813	0.0409	0.1797

As before, the estimates were compared using a scatter plot. This is shown in Figure 2.

Figure 2: Scatter plot of small area estimate of poor health and Census data



3 Calculation of ICCs

Intra-class correlation coefficients were generated for both outcomes. These are Wales-wide values - unknown population quantities that can be estimated from both the survey data and modelled data.

The direct estimate and the estimate based on the small area estimates are both estimates of the same quantity and there is no issue in comparing them. This comparison is made as an additional validation check on the quality of the model, so this is not a circular approach. The idea is that the direct estimate is more reliable as it does not involve making any modelling assumptions. Large differences between the two estimates suggest there is an error in the model.

ICC calculated from the small area estimate

Suppose the population consists of a large number of small areas (such as MSOAs), which, for simplicity I am assuming are of equal size. The outcomes are binary outcome (such as being a Welsh speaker or not) and the aim is to calculate means rather than doing any logistic or probit regressions.

Let p_i be the proportion of people in the *i*-th area who speak Welsh and let p be the overall mean. The inter-cluster correlation (see Appendix C) can then be calculated as:

$$\frac{\mathrm{V}(P)}{E(P)\big(1-E(P)\big)}.$$

The modelled proportions are used to calculate the variance and expected value. These are entered into the equation above to estimate the ICC.

A second, direct, value of rho can be generated by entering the Census mean and variance into the equation.

A third estimate can be obtained from the National Survey data using Stata's **loneway** command.

If these three estimates are not approximately equal then there must be doubts about the model and the small area estimates.

3.1 Examples

The small area estimates for Welsh speaking have an average value of 0.1913 and variance of 0.0280. Using the formula above (and assuming the MSOAs are of equal size) the ICC would equal 0.0280/(0.1913*0.8197) = 0.181.

This calculation was repeated using the population mean and variance of the Census data. These show Welsh speaking to have an average value of 0.1908 and variance of 0.0289, which is an ICC of 0.187.

Estimating this value directly from the National Survey data using the **loneway** command gives an ICC of 0.241.

These three values are relatively close. A final check was made using the xtlogit command with the National Survey data and an empty model. This gave a value of rho of 0.328. However, the rho calculated by xtlogit represents a different measure from that calculated by loneway; xtlogit treats the data as binary and loneway as numerical. In addition, xtlogit does not require the MSOAs to be of equal size. Hence we would expect some difference and should not be too concerned.

The corresponding values for poor health were 0.009 from the small area estimates, 0.009 using Census data and 0.010 for the direct estimate using the National Survey. The value generated using xtlogit was 0.035, again higher due to the different data assumptions behind the calculation.

The values for ICC are generally close, with the exception of the value generated using xtlogit. The figures above imply that there is much more variation across MSOAs in Welsh speaking than poor health.

Appendix A: Stata code for outcomes

```
lab def YESNO 1 "Yes" 0 "No"


gen PoorHealth = inlist(GenHealth, 4, 5) if inrange(GenHealth, 1, 5)
lab var PoorHealth "Poor health"
lab val PoorHealth YESNO
tab GenHealth PoorHealth, miss


gen SpkWelsh = (DvWelSpk == 1) if inlist(DvWelSpk, 1, 2)
lab var SpkWelsh "Speaks at least a little Welsh"
lab val SpkWelsh YESNO
tab DvWelSpk SpkWelsh, miss
```

Appendix B: Covariates

Variable name	Variable label
Census2011	
male	% usual residents - male
pop0_17	% usual residents - 0-17 yrs
pop18_64	% usual residents - 18-64 yrs
pop18_34	% usual residents - 18-34 yrs
pop35_64	% usual residents - 35-64 yrs
pop65plus	% usual residents - 65+ yrs
single	% usual residents aged 16+ - single
married	% usual residents aged 16+ - married
nonwhite	% usual residents - from a non white ethnic group
bornWales	% usual residents - born in Wales
bornRestUK	% usual residents - born in UK (not Wales)
bornNotUK	% usual residents - born outside UK
noreligion	% usual residents - no religion
workALL	% usual residents aged 16-74 - in work (including self-employed)
workFT	% usual residents aged 16-74 - full time employees
workPT	% usual residents aged 16-74 - part time employees
manprof	% usual residents aged 16-77 in employment - mangerial or professional occupations
llti	% usual residents - limiting long term illness
qual4	% usual residents aged 16+ - highest qualification is level 4
noqual	% usual residents aged 16+ - no formal qualifications
nssecMan	% usual residents aged 16-74 - managerial NSSEC group
ownpers	% usual residents in households - owner occupiers
socpers	% usual residents in households - social renting
privpers	% usual residents in households - private renting
ownhhds	% households - owner occupiers
sochhds	% households - social renting
privhhds	% households - private renting
detached	% household spaces - detached
flats	% household spaces - flats

hh1pers	% household spaces with 1 usual resident - with one person
depchildn	% families in households - all dependent children
depyoung	% families in households - youngest dependent child aged <5yrs
nocar	% households - with no car
gradeA	HRPs - social grade A
popdens	Population density (population in private households / area of the MSOA)
ACORN	
DvACORN	Derived variable - ACORN classification (LSOA)
pacorn1	% households in the MSOA belonging to ACORN category 1
pacorn2	% households in the MSOA belonging to ACORN category 2
pacorn3	% households in the MSOA belonging to ACORN category 3
pacorn4	% households in the MSOA belonging to ACORN category 4
pacorn5	% households in the MSOA belonging to ACORN category 5
pacorn6	% households in the MSOA belonging to ACORN category 6
Welsh IMD	
pimd15	% LSOAs in the MSOA belonging to most deprived IMD quintile
DvWIMDEdu5	Derived variable - Welsh Index of Multiple Deprivation education score (quintiles)
DvWIMDEmp5	Derived variable - Welsh Index of Multiple Deprivation employment score (quintiles)
DvWIMDEnv5	Derived variable - Welsh Index of Multiple Deprivation physical environment score (quintiles)
DvWIMDHlth5	Derived variable - Welsh Index of Multiple Deprivation health score (quintiles)
DvWIMDHse5	Derived variable - Welsh Index of Multiple Deprivation housing score (quintiles)
DvWIMDInc5	Derived variable - Welsh Index of Multiple Deprivation income score (quintiles)
DvWIMDOvr5	Derived variable - Welsh Index of Multiple Deprivation overall score (quintiles)
DvWIMDSafe5	Derived variable - Welsh Index of Multiple Deprivation community safety score (quintiles)
DvWIMDServ5	Derived variable - Welsh Index of Multiple Deprivation access to services score (quintiles)
Other areas	
DvAsEcArea	Derived variable - Assembly Economic Fora Area
DvComFirst	Derived variable - Communities First areas (binary)

There were a large number of potential covariates, many of which were highly correlated. Factor Analysis and correlations were used to reduce these to a more manageable number. As a result, the following variables were dropped from the analysis:

- Household-level versions of owners, private renters and social renters (ownhhds, sochhds and privhlds) plus the individual-level version of social renters (socpers)

- Variables on non-manual occupations and social grade (nssecMan and gradeA) as they were correlated with the qualification variables

- Variables on country of birth - bornRestUK and bornNotUK - as they were correlated with bornWales and non-white,

- The proportion of families with young children (depyoung) as it was correlated with proportion of families with dependent children (depchildn),

- Male as it did not vary much across areas.

Appendix C: Models

The full output from the models used to generate the small area estimates is given below.

Welsh speaking

Number of observations = 14657

Number of groups = 410

Obs per group: min = 3

average = 35.7

max = 117

Integration points = 12

Wald chi2(33) = 1109.03

Prob > chi2 = 0.0000

Log likelihood = -5896.6095

					95% Confidence Interval	
SpkWelsh	Coef.	Std.Err.	z	P>z	Lower	Upper
DvUniAuth: Local Authority						
Anglesey						Baseline
Gwynedd	0.242	0.214	1.1	0.260	-0.179	0.662
Conwy	-0.619	0.229	-2.7	0.007	-1.067	-0.170
Denbighshire	-0.902	0.223	-4.0	0.000	-1.339	-0.465
Flintshire	-1.431	0.242	-5.9	0.000	-1.906	-0.956
Wrexham	-2.047	0.229	-8.9	0.000	-2.497	-1.598
Powys	-0.835	0.706	-1.2	0.237	-2.219	0.548

Ceredigion	0.239	0.758	0.3	0.753	-1.247	1.725
Pembrokeshire	0.842	0.831	1.0	0.311	-0.788	2.471
Carmarthenshire	2.534	0.908	2.8	0.005	0.755	4.313
Swansea	1.245	0.935	1.3	0.183	-0.588	3.078
Neath Port Talbot	1.450	0.999	1.5	0.147	-0.508	3.409
Bridgend	-0.302	0.887	-0.3	0.733	-2.042	1.437
Vale of Glamorgan	-0.109	0.844	-0.1	0.897	-1.763	1.545
Cardiff	0.377	0.840	0.5	0.654	-1.270	2.024
Rhondda, Cynon, Taf	-0.203	0.920	-0.2	0.826	-2.006	1.601
Merthyr Tydfil	0.004	0.942	0.0	0.997	-1.842	1.849
Caerphilly	-0.301	0.924	-0.3	0.744	-2.112	1.510
Blaenau Gwent	-0.799	0.941	-0.9	0.396	-2.644	1.045
Torfaen	-0.729	0.906	-0.8	0.421	-2.503	1.046
Monmouthshire	-0.626	0.749	-0.8	0.403	-2.095	0.842
Newport	-0.902	0.875	-1.0	0.303	-2.617	0.813
nocar	-0.042	0.012	-3.5	0.000	-0.066	-0.019
bornWales	0.050	0.006	8.1	0.000	0.038	0.062
married	-0.026	0.011	-2.3	0.022	-0.048	-0.004
hh1pers	0.028	0.012	2.3	0.024	0.004	0.052
qual4	0.001	0.008	0.1	0.931	-0.015	0.016
DvWIMDSafe5: IMD safety quintiles						
Q1 Most deprived 20%						Baseline
Q2	0.330	0.100	3.3	0.001	0.134	0.525
Q3	0.260	0.102	2.5	0.011	0.059	0.460
Q4	0.404	0.107	3.8	0.000	0.194	0.615
Q5 Least deprived 20%	0.608	0.116	5.2	0.000	0.380	0.836

DvAsEcArea						
Mid Wales	0.000					
South West Wales	0.000					
South East Wales	0.000					
bornWales	0.000					
Interaction of DvAsEcArea and bornWales						
Mid Wales	-0.012	0.012	-1.0	0.319	-0.036	0.012
South West Wales	-0.049	0.012	-4.0	0.000	-0.072	-0.025
South East Wales	-0.038	0.011	-3.5	0.001	-0.060	-0.016
Constant	-2.117	0.889	-2.4	0.017	-3.860	-0.374
/lnsig2u	-1.725	0.183			-2.083	-1.367
sigma_u	0.422	0.039			0.353	0.505
rho	0.051	0.009			0.036	0.072

Likelihood-ratio test of rho=0: chibar2(01) = 77.02 Prob >= chibar2 = 0.000

Poor Health

Number of observations = 14652

Number of groups = 410

Obs per group: min = 3

average = 35.7

max = 117

Integration points = 12

Wald chi2(4) = 112.54

Prob > chi2 = 0.0000

Log likelihood = -4675.5326

					95% Confidence Interval	
Poorhealth	Coef.	Std.Err.	z	P>z	Lower	Upper
Communities First Areas						
(1 April 2013)	0.229	0.071	3.2	0.001	0.089	0.369
pacorn1	-0.005	0.002	-2.2	0.030	-0.009	0.000
llti	0.038	0.008	4.6	0.000	0.022	0.054
depchildn	0.019	0.007	2.8	0.005	0.006	0.032
Constant	-3.872	0.397	-9.8	0.000	-4.650	-3.095
/lnsig2u	-3.693	0.886			-5.429	-1.957
sigma_u	0.158	0.070			0.066	0.376
rho	0.008	0.007			0.001	0.041

Likelihood-ratio test of rho=0: chibar2(01) = 1.54 Prob >= chibar2 = 0.107

The output includes the additional panel-level variance component. This is parameterized as the log of the variance (labelled lnsig2u in the output). The standard deviation is also included in the output and labelled sigma u together with rho. Rho is the proportion of the total variance contributed by the panel-level variance component.

Rho is small, particularly for the poor health model, this is because the variables in this model are MSOA-level. Whilst we had the option of including age and sex of the individual, these variables were not the strongest predictors and were left out of the final model.

Appendix D: Generating the ICC

The ICC can be calculated by comparing a simple random sample of size two, with a clustered sample of size two (one cluster chosen at random two individuals chosen from that cluster). In that case, the ratio of the variance of the estimators will be 1 + ICC.

Simple random sampling

Taking a simple random sample of size two gives an estimate which is unbiased with a variance equal to $p(1-p)/2$.

Clustered sample

If we choose one cluster at random and then choose two individuals in that cluster the estimate, T, is also unbiased. We can use the law of conditional expectation to calculate its variance. Conditional on having chosen cluster j, the estimate has a mean of p_j and a variance of $p_j(1-p_j)/2$. So, the variance of T is:

$$V(T) = V\{E(T|j)\} + E\{V(T|j)\} = V(p_j) + E\left(\frac{p_j(1-p_j)}{2}\right)$$

Simplify this:

$$V(p_j) = E(p_j^2) - p^2,$$

and

$$E\left(\frac{p_j(1-p_j)}{2}\right) = \frac{p}{2} - E\left(\frac{p_j^2}{2}\right).$$

So

$$V(T) = E\left(\frac{p_j^2}{2}\right) + \frac{p}{2} - p^2 = E\left(\frac{p_j^2}{2}\right) - \frac{p^2}{2} - \frac{p^2}{2} + \frac{p}{2} == E\left(\frac{p_j^2}{2}\right) - \frac{p^2}{2} + \left(\frac{p(1-p)}{2}\right).$$

Divide the variance of the clustered estimator by that of the simple random sample estimator ($p(1-p)/2$) and we obtain:

$$\text{Ratio} = 1 + \frac{E\left(\frac{p_j^2}{2}\right) - \frac{p^2}{2}}{\frac{p(1-p)}{2}} = 1 + \frac{E(p_j^2) - p^2}{p(1-p)} = 1 + \frac{V(p_j)}{p(1-p)}.$$

As the ratio is 1 + ICC, this gives the ICC.