Session Number: Parallel Session 4D
Time: Tuesday, August 24, PM

Paper Prepared for the 31st General Conference of
The International Association for Research in Income and Wealth

St. Gallen, Switzerland, August 22-28, 2010

Scale Issues in the Analysis of Spatial Variations in the Distribution of Household Income: Developments in Data Linkage in a New Longitudinal Study

Nick Buck

For additional information please contact:

Name: Nick Buck
Affiliation: University of Essex

Email Address: nhb@essex.ac.uk

Scale issues in the analysis of spatial variations in the distribution of household income: developments in data linkage in a new longitudinal study

Nick Buck[1]
Institute for Social and Economic Research, University of Essex

1. Introduction

As the presence of this session of the IARIW conference indicates, linkage between surveys and external data is of increasing interest. As we note below, there are various types of linkage based on different types of external sources, While the majority of papers in this conference stream concern combination of survey and administrative data, linked at the individual level, another important type concerns combination of survey data with spatially referenced data. The role of this paper is to argue for the value of this approach to data linkage. Through some examples based on linkage of the first wave of the *Understanding Society*, the new UK Household Longitudinal Study (UKHLS), it provides evidence of the value of such linkage for the analysis of inequalities between households, and specifically income inequalities. Because the analysis focuses on the first wave of a new longitudinal survey, the results presented here are necessarily cross-sectional, but we anticipate some research which might be done in later waves when longitudinal data become available. Both spatial and administrative data linkage are key objectives for the UKHLS, and the paper includes a brief overview of the overall plans. However its major focus is on spatial linkage and it provides early analysis of the first wave of the survey.

There is an increasing interest in quantitative social sciences that the spatial context within which social or economic processes take place may have significant effects. Place is of course only one of a range of possible types of context. Others include institutional contexts such as school or workplace, or indeed other social processes by which reference groups might be formed. There is, however, particular interest in place as a context, in part because of developments in social policy which have again highlighted the role of spatial policy, and a related perception that space may be a significant dimension in structuring social and economic inequality. Because some sorts of information about place and local context may not be directly available to survey participants, especially for example local quantitative indicators of population composition and behaviour, there is natural role for linkage to make these data available to researchers. One of the core goals in establishing UKHLS was to extend the range of measures available to analysts, and one key direction was the context for participants behaviour – hence the key role for spatial data linkage.

The research and policy interest in spatial dimensions to the distribution of income, deprivation and welfare relates both to issues of population composition and of potential contextual effects. The former concerns segregation by income or deprivation at different spatial scales which may be seen as problematic in its own

[1] The research was supported by ESRC funding for *Understanding Society* (RES-586-47-0001) and the ESRC UK Longitudinal Studies Centre. Understanding Society is an initiative by the Economic and Social Research Council, with scientific leadership by the Institute for Social and EconomicResearch (ISER), University of Essex and survey delivery by the National Centre for Social Research (NatCen). Nick Buck is Principal Investigator for the Understanding Society.

right or indeed as an opportunity to guide the spatial targeting of policy interventions. The latter is based on the hypothesis that there are individual or group effects arising from uneven spatial distribution: that, for example, having poorer neighbours may have negative effects on individuals' life chances. There is a range of potential policy implications from compositional differences and neighbourhood effects. The former may lead to spatial targeting of redistributive policy, while the latter may motivate attempts to ameliorate negative neighbourhood effects.

A range of different approaches to the analysis of spatial data in combination with survey data with different overall goals follow in part from the distinction between area variation arising purely from compositional differences and variations arising from contextual effects. In the first instance it is not hypothesised that living in one area rather than another makes any particular difference to outcomes or behaviour for individuals. In the second case there is a hypothesis that some aspect of the area context has such an impact on outcome or behaviour. Clearly the latter is a much more powerful motivation for data combination since the combination involves the addition of information hypothesised to have an independent influence on outcomes of interest.

However there are some significant identification problems associated with such models. These are largely beyond the scope of this paper (see Durlauf (2004) and Dietz (2002) for discussion. The issues are most acute where the interest is in the analysis of social interactions, i.e. where individual behaviour is hypothesised to be influenced by group (e.g. other area residents) behaviour (see Manski 1993 and 2000). Issues may be somewhat more tractable where the feature of the area context of interest is more clearly exogenous to area population characteristics, e.g. environmental hazards or policy interventions, though even here there may be associations, e.g. because of policy targeting.

However understanding the extent of spatial variation and how it relates to population composition, via association with population characteristics still remain valuable. The spatial structuring of inequality is a subject of interest in itself, both how intense those inequalities are and at what spatial scale they operate. This provides evidence on the sources of such inequalities, and thus on the potential policy interventions to mitigate them. For example, inequalities manifested at a larger spatial scale, close to the regional scale, might be thought to relate to relatively entrenched differences in the pattern of economic opportunities, whilst smaller scale inequalities might be hypothesised to relate to differences processes leading to residential segregation. This paper does not address the question of the sources of spatial variation. Rather it aims to suggest ways in which analysis of surveys in combination with spatial data might be able to contribute to such analysis.

This paper focuses on the analysis of variations in household income. This is clearly only one of a wide range of potential indicators of spatial variation or spatial inequalities. One could alternatively for example examine variations in poverty or low income risk or other measures of deprivation or economic well-being. It should be noted that in the UK, as in many other countries, we do not have population level measures of income for small areas derived from administrative sources. The closest is the Survey of Personal Incomes derived from a sample of tax records, but this is not used to produce small area estimates. In some countries there is data based on commercial marketing surveys (see for example the data used by Knies el al (2008) for Germany). Such information is not yet available in the UK. The data combination

approaches used here are distinct from, but relate to the approach to use survey data to derive local small area estimates (e.g. Rao (2003), Pfefferman (2002)).

One key issue which arises in spatial linkage, which does not arise in quite the same way for other types of linkage is a spatial scale issue. Spatial linkage normally involves defining a territorial unit which includes the participant[2]. Whereas for administrative or organisational linkage, the unit to be linked is normally clear, for spatial linkage there will be a wide range of choices. This may have major substantive implications for modelling of associations and effects, and the choices need to be well motivated. This is a major theme of the analysis presented in this paper and is discussed further below.

In the rest of this paper we outline some of the issues involved in linking surveys to external data, discuss the approach to investigating different spatial scales, and present some more information on the data being used from the UKHLS. We then present two analyses by way of example of the role of survey and spatial data combination. The first uses income inequality decomposition approaches to explore the extent of local area inequality and how it contributes of overall inequality. The second explores the associations between individual household income levels and local area characteristics at a range of spatial scales.

2 Linkage of surveys to external data

The ability to link data collected directly from survey participants with other data sources would enhance the scientific research capacity of the survey substantially. Data linkage can be used to provide:

(i) supplementary data that could not economically or reliably be collected in the survey, extending the coverage of topic areas – quantitative spatial data often falls into this category;

(ii) substitute data for information that could otherwise be collected in the survey (thereby potentially reducing participant burden and freeing up space for other questions), and

(iii) a means of validating survey data when the same sorts of information are available in the survey and the other source(s). More generally, administrative record data are commonly argued to be more accurate than survey data, and so their use may improve survey quality.

The types of data that might be linked in to surveys may be classified in several ways. First, there are data for which the information is at the level of the individual, e.g. social security benefits received, or an event such as a spell in hospital. Second, there are data at the level of an organisation, e.g. the characteristics of the school attended by a sample member, or of the employer for whom a participant works. Third, there are data which refer to a unit that is defined spatially, e.g. a neighbourhood or some other geographically-defined area such as a local labour market.

Full population data sets, including both censuses and administrative data sets provide the potential for information at small spatial scales, and area level indicators, e.g. deprivation indicators, are often constructed from such data. However, compared with household surveys which focus on income and welfare issues they often contain

[2] There are other potential forms of spatially related linkage which do not exactly have this character. For example linkage may also involve defining measures of distances, e.g. from public service facilities, environmental hazards, the location of other family members.

rather weak information on individual outcomes and behaviours and in many cases rather poor information on individual and household measures of income and well-being. In many countries also there is limited access to unit record data from population data sets. There is thus a case for matching surveys with spatial data from population data sets in order to explore the spatial dimension to the distribution of income.

There is a wide range of potential spatially referenced data sources which could be linked, and in the UK the range has expanded rapidly in recent years. These include Census Small Area Statistics, a wide range of data derived from administrative sources, including labour market data from employment and unemployment registrations, benefit receipt data, reported crime, sources based on transactions data, including house price measures, but also potential new sources based on information from retailing and electronic communications. There have been a number of geo-demographic profiling systems based on these data. Other potential sources include data about public expenditures and the delivery of public services and data on the quality of the physical environment, e.g. air pollution or measures of the amount of green space derived from remote sensing. While there are significant costs associated with assembling some of these data sets, many of them are now publicly available in ways which relatively straightforward to link to survey data provided there is access to address information. This does mean that there is a particular set of disclosure risks associated with spatial data linkage. This means that the linkage operations will need to be made by the study organisation holding the addresses under secure conditions, and linked data made available to researchers under 'safe' access conditions, either through additional data access agreements or in safe settings.

In the UK the methodology for linking spatial data to addresses is relatively straightforward. The address includes a seven character postcode, which for residential addresses covers around 10 dwellings. Tthese can be linked via the National Statistics Postcode Directory both to a range of administrative and statistical geographies, and to the grid reference, a high resolution location indicator, which for most postcodes will be accurate to the nearest metre for the location of the building closest to the postcode mean.

Spatial scale issues

One of the key issues in the analysis of combined data sets including surveys and spatial data is the spatial scale of matching. There are various substantive reasons for choosing particular scales: appropriate units of policy intervention, areas which are perceived to be socially or economically meaningful to individuals, or areas which prove empirically to be most effective in capturing area effects. In practice however the choice of spatial scale for matching tends to be pragmatic, based on the consistent scale indicators which are available in the data sets being matched. These areas may not be particularly meaningful either for policy purpose or in relation to individual behaviours. For example there may be considerable internal heterogeneity, or individuals may be located towards the boundary of an area, so that even if the scale is relevant to their behaviour the particular area of the standard geography within which the fall may not exclude large parts of the area which is relevant for their behaviour.

This paper focuses in part on testing of the appropriate scale for matching. It uses a range of standard pre-defined geographies and also a more flexible approach which allows the definition of areas at other scales. It is based on 'bespoke areas'

which rely on having exact survey participant locations and aggregating very small areas from e.g. census geographies based on proximity to participant location to create areas at a range of scales with the participant located close to the centre. This overcomes some of the problems above, and allows an empirical test of relevant spatial scale. It cannot deal with the idiosyncrasies of particular locations, for example where a physical boundary limits movement.

The bespoke areas were created by aggregating census Output Areas (with an average population of just over 250 people) closest to the participant location, until a range of threshold sizes are reached. There are two versions, one based on aggregating to reach threshold population sizes or aggregating to include all Output Areas whose centroid lies within a range of distances from the postcode centroid. The population size bands used here are 500, 1000, 2000, 5000 and 10,000 people. The distance bands used are 250 metres, 500 metres, 1,000 metres and 2,000 metres. In sparsely populated there may be some households where no Output Area centroid falls within the shorter distance of residence. It would clearly be possible to substitute the Output Area of residence in these cases, but here they have been excluded from the analysis.

In addition to these bespoke areas we use a number of more standard geographies. These include the Census Output Area introduced above and the Postcode Sector which was used for sampling purposes. The mean and standard deviations of population sizes of these units are shown in Table 1 below. In addition we use two large geographies. The local authority district is the administrative unit for many purposes. There is a rather large range of sizes of these units, and many districts are quite heterogeneous. For an alternative geography at a large scale we also use parliamentary constituencies, which are more uniform in size and somewhat more homogeneous. Again, population details are given in Table 1.

Table 1: Population of units used in analysis (2001 Census)

	Mean	std dev
Census output area	262	98
Postcode Sector	5978	3698
Parliamentary Constituency	90898	11095
Local Authority District	139960	94500

Data: the UK Household Longitudinal Study

UKHLS, also known as *Understanding Society,* is a study which follows individuals over time, regularly collecting data about each sample member and his or her household. Such household panel studies have provided a major resource for understanding key issues which face societies around the world. They provide unique information on the persistence of such states as child poverty or disability, on factors which influence key life transitions, such as marriage and divorce, and they provide information on the effects of earlier life circumstances on later outcomes. They also support research relevant to the formation and evaluation of policy and enable improved and more reliable analytical techniques which cross-sectional data, based on only a single observation of each individual, cannot support. In the UK, the British Household Panel Survey has been particularly successful, has already been accessed by more than 2000 users and generates more than 150 publications per year; it is heavily used by government departments and by researchers outside the UK. UKHLS

has been built on the success of the BHPS, and represents a major advance on it. For further information on the UKHLS please see http://www.understandingsociety.org.uk.

In particular it will have a very much larger sample size. The target sample of 40,000 households will give a unique opportunity to explore issues for which other longitudinal surveys are too small to support effective research. It will permit analysis of small subgroups, such as teenage parents or disabled people. Examples include analysis at regional and sub-regional levels, allowing examination of the effects of geographical variation in policy (notably differences between the countries of the UK). A large sample size also allows high-resolution analysis of events in time, for example focussing on single-year age cohorts. The sample at the start is representative of the whole UK population, and with properly designed following rules will remain representative of that population, subject to adjusting for attrition.

At each wave all household members are interviewed. This has major advantages for important research areas such as consumption and income, where within-household sharing of resources is important, or demographic change, where the household itself is often the object of study.

Other areas of innovation of innovation for the UKHLS include a major strand of data collection and questionnaire design to support research into ethnic minority groups, the collection of biomarkers and other health indicators, and an extensive programme of collection of linked data, noted above.

The questionnaire involves rather broad, interdisciplinary topic coverage, going beyond that traditional included in household panel studies. It does however include key measures of household economic well-being including all major components of individual and household income.

The UKHLS sample of 40,000 households will include:

a) An Innovation Panel of 1500 households to enable methodological research, including experiments for developing and assessing mixed modes of data collection. The fieldwork for the Innovation Panel commenced in January 2008.
b) A new equal probability general population sample achieved sample of around 27,000 households. The fieldwork for this sample commenced in January 2009.
c) A boost ethnic minority sample, of around 4,000 households to provide 1,000 adult individuals in each of five minority ethnic groups: Caribbean, African, Indian, Pakistani, and Bangladeshi, plus members of other ethnic minorities in the areas covered by the boost sample, to supplement the ethnic minority respondents in the main sample. The fieldwork for this sample will run alongside that for the general population sample.
d) The British Household Panel Survey (BHPS) sample of approximately 8,400 households, from which data has been collected since 1991. Wave 18 of BHPS will be collected at the same time as wave 1 of UKHLS, and the sample will be integrated into the new study from wave 2, starting in January 2010.

The analysis in this paper is based on the first year of the new general population sample. This is based upon a stratified, clustered, equal probability sample of residential addresses drawn to a uniform design throughout the whole of the UK (including north of the Caledonian Canal), with the exception of Northern Ireland. In Northern Ireland, the sample is unclustered.

Primary Sampling Units (in GB) are postal sectors. Postal sectors will first be stratified by an appropriate set of Census variables and then 2,640 selected systematically with probability proportional to number of addresses. The stratification variables were selected on the grounds of likely correlation with key survey measures and to include a regional indicator plus two or three other socio-economic indicators. Within each sampled sector, 18 PAF addresses were selected, resulting in an equal-probability sample of a total of 47,520 addresses in GB. In NI, 2,400 addresses will be selected systematically from the Land and Property Services Agency list of domestic properties, thus making a total of 49,920 selected addresses in UK.

Understanding Society is based on twelve month intervals between interviews for each respondent. Fieldwork for each wave is continuous over a 24 month period, with each monthly sample nationally representative. . In this design the second wave fieldwork overlaps with that for the first wave so as to preserve the twelve month interval between individual interviews. This arises from fieldwork capacity constraints, especially in the early waves. All sample members aged 16 and over are eligible for the adult interview, and there is an additional questionnaire for children aged 10-15.

The UKHLS therefore has particular advantages for spatial analysis, following from its large sample size and the capacity to undertake spatial data linkage. At wave of the survey we are dealing with a nationally representative random sample, clustered in ways which permit the analysis of variation within the sampling clusters. In later there will be residential mobility and hence declustering. However the processes of spatial mobility are themselves of substantive interest, and are relevant for example for behaviour relevant to the evolution of household income.

This paper is based on the first year of fieldwork for wave one, carried out between January and December 2009. This provides a national sample of around 14,000 households. The analysis is based on a very early pre-release version of the data, meaning that the results presented here need to be regarded as provisional. In particular, the analyses are unweighted, and the household income measure used in gross income, including benefits, but before any deduction of taxes. Derived net income variables for the study will be produced in due course. The income data includes imputation for missing data. The models used are subject to further development.

Given differences in the sampling schemes used the analysis here is restricted to Great Britain.

The spatial component of income inequality

In this section we explore how far there is variation in income between different spatial units. As Berthoud (2008) observes, there is a common assumption of differences in household incomes between places, but as he also observes, until we get to very small areas this variation does ‘explain’ high proportion of overall income distribution. In that paper he uses an analysis of variance approach. Here we use a decomposition of inequality indices to address the same issue, using the UKHLS wave 1 year 1 data.

We focus on decomposable inequality indices, i.e. indices for which it is possible to additively decompose total inequality into a part representing between population sub-groups and a part representing inequality within sub-groups in a way

which are consistent given any change in inequality. It has been shown that the main class of indices for which this is possible are the generalised entropy measures (Shorrocks 1984, Cowell 1985). Other perhaps more familiar indices such as the Gini coefficient cannot be decomposed consistently in this way. Three indices from this group are used, varying in their sensitivity to inequalities at different points in the income distribution (for an application see Mookherjee and Shorrocks, 1982). These are (with formula based on unweighted data),

The Mean log deviation:

$$I_0 = \frac{1}{n}\sum_i \log\left(\frac{\mu}{y_i}\right)$$

The Theil index:

$$I_1 = \frac{1}{n}\sum_i \frac{y_i}{\mu} log \frac{y_i}{\mu}$$

and ½ coefficient of variation squared:

$$I_2 = \frac{1}{2n\mu^2}\sum_i (y_i - \mu)^2$$

We can additively decompose these indices into those parts which reflect the inequality within each group. Here v_k is the proportion of the population in the k-th group and $\lambda_k = \mu_k/\mu$ is its mean income relative to the whole population.

Thus for the Mean Log Deviation we have:

$$I_0 = \sum_k v_k I_0^k + \sum_k v_k log(1/\lambda_k)$$

for the Theil Index:

$$I_1 = \sum_k v_k \lambda_k I_1^k + \sum_k v_k \lambda_k log \lambda_k$$

and for ½ CV^2 :

$$I_2 = \sum_k v_k (\lambda_k)^2 I_2^k + \frac{1}{2}\sum_k v_k [(\lambda_k)^2 - 1]$$

The first term in each equation is the within-group component, a weighted sum of the subgroup inequality components. The second term is the between group component based on the contribution of inequalities in subgroup means.[3]

It should be noted that the bespoke area approach is not straightforwardly usable for inequality decomposition – since this depends on having multiple households within each territorial unit, and the bespoke areas are defined separately for each household. We focus here on standard geographies, including the geography used in sampling, the postcode sector, and secondly two larger scale geographies, Parliamentary Constituencies and Local Authority Districts, discussed above. Table 2 below shows the distribution of sample households across these units. It should be noted that these distributions are very different, reflecting the different scales. So, in

[3] Analysis was undertaken using Stephen Jenkins's Stata programme *ineqdeco* (see Stata Technical Bulletin 48, 4-18, 1999)

terms of numbers of sample households, the median postcode sector size is 9, the median constituency size is 28, and the median district size is 42.

Table 2: Distribution of UKHLS year 1 sample by spatial units

	Postcode sector		Constituency		LA district	
Number of households per area unit	N areas	N households	N areas	N households	N areas	N households
1-5	211	677	26	78	12	27
6-10	814	6806	78	654	39	319
11-15	526	6414	102	1289	52	653
16-20	12	196	86	1545	40	720
21-30			163	4187	80	2046
31-40			81	2814	62	2190
41-50			33	1457	55	2465
50 & over			35	2234	79	5660

In the analysis presented in table 3 below, we decompose gross house income equivalised using the modified OECD equivalence scale. We exclude zero and negative incomes, as well as a very small number of outliers at the top end of the distribution.

Table 3: Decomposition analysis

	Mean log deviation	Theil index	½ CV squared
Total index	0.25754	0.29776	0.66139
Spatial groupings			
Region			
between group	0.00601	0.00611	0.00624
% of total	2.3%	2.1%	0.9%
District			
between group	0.02846	0.03057	0.03453
% of total	11.1	10.3	5.2
Constituency			
between group	0.03859	0.04088	0.04601
% of total	15.0	13.7	7.0
Postcode sector			
between group	0.07124	0.08139	0.10749
% of total	27.7	27.3	16.3
Non spatial groupings			
Household type			
between group	0.02426	0.02267	0.02172
% of total	9.4	7.6	3.3
Housing tenure			
between group	0.02661	0.02416	0.02241
% of total	10.3	8.1	3.4
Number in employment			
between group	0.03628	0.0338	0.0321
% of total	14.1	11.4	4.9
Household type & number in employment			
between group	0.05842	0.05288	0.05049
% of total	22.7	17.8	7.6
Housing tenure & number in employment			
between group	0.05188	0.04691	0.0439
% of total	20.1	15.8	6.6
Housing tenure & household type			
between group	0.04577	0.04223	0.04062
% of total	17.8	14.2	6.1
Household type, housing tenure & number in employment			
between group	0.07134	0.06443	0.06201
% of total	27.7	21.6	9.4

One can look at the evidence from this analysis in three different ways. Firstly, how far does the between-group component increase as we move to a progressively more disaggregated geography? The evidence here is quite clear. While the region based grouping accounts for only around 2% of the mean log deviation indicator, the district level accounts for 11% and the constituency level accounts for 15%. It does appear that the slightly smaller size, and rather greater homogeneity of constituencies compared with local authority districts means that the former grouping is substantially more effective at accounting for inequality differences. The between group component at postcode sector level accounts for more than 27% of the mean log deviation based inequality. There is clearly a question here for further work of how far this is being driven by the relatively larger number of areas, including some with very small numbers of households.

Secondly, there is the question of the comparison of the spatial groups with household based groupings. This is introduced essentially to give some intuition of the substantive importance of the spatial groupings. The three single level groupings, household type (7 categories), housing tenure (5 categories) and number in employment (4 categories) each account for around the same proportion of income inequality as the district and constituencies levels. Given that the household based groupings use many fewer categories than the spatial groupings it is clear they are on the whole a more parsimonious way of accounting for income inequality. However it does suggest that spatial categories have equivalent levels of effectiveness as factors at the household level which are well know to have a direct association with household income, even after income is equivalised. When the household level characteristics are combined, so as use their interactions as groupings, the between group component rises substantially to levels above those for district or constituency groupings. However groups based on the three-way interaction of household type, housing tenure and number in employment accounts for around the same proportion of inequality as the postcode sector level.

Thirdly we can compare the different inequality indicators. As indicated above, they differ in their sensitivity to inequalities in different parts of the income distribution, with the mean log deviation most sensitive to inequalities affecting the lower end of the distribution, and the measure based on $1/2\ CV^2$ most sensitive to inequalities affecting the upper of the distribution. The between group share for all groupings falls consistently as we move towards indicators more sensitive to inequalities in higher incomes, suggesting the groupings are more effective at partitioning those in the lower part of the distribution. However, between group component of the$1/2\ CV^2$ measure does appear relatively higher for spatial measures compared with the household level measures, suggesting that somewhat more variation in higher incomes is being captured in the geographical segmentation.

The association between household income and area characteristics

In this section we present some models of the association between household income and area characteristics at a number of different spatial scales, both controlling and not controlling for household characteristics. These models are rather exploratory. As indicated above, the analysis is based on rather preliminary version of the UKHLS data, and one could develop the range of household level predictors included in the models as well as working with different income concepts. Moreover analysis is somewhat illustrative, and it is certainly not being used to test a hypothesis of a neighbourhood context effect on income levels. There are however a range of

potential extensions to this work, in which it would be possible to explore hypotheses about the association between area characteristics and income and other related measures (e.g. poverty and low income risk) as well as associations with changes in these measures in a longitudinal context (e.g. Bolster et. al. 2007).

Here we define a very simple OLS regression model for predicting log equivalised household income including both household and area characteristics. Table 4 shows the model with household characteristics only. They include two housing tenure categories, along with measures of the household structure, and number of workers. It would clearly be possible to derive a fuller model.

Table 4: Regression model for log equivalised household income: household level characteristics

	Coef.	t
Social housing renter	-0.21986	-12.87
Own housing with mortgage	0.219596	14.83
Number in employment in household	0.364303	48.26
Number of children in household	-0.12502	-17.16
Number aged 65 & over in household	0.073509	6.52
Household consisting of lone parent family	-0.06797	-2.71
Constant	6.816649	467.31
R squared	0.2885	

Our predictor variables at the area level have been selected from the 2001 Census of Population Small Area Statistics. There are issues in select area level variables, and especially in identifying plausible predictors. This is an area for significant further work[4]. It would useful to incorporate more up to date indicators, e.g. of labour market situation in the current recession, though in the UK at least there is considerable persistence in the spatial distribution of area characteristics, and in general, areas which experienced particularly high unemployment in the current recession are likely to be the same as those with high unemployment eight years earlier. In the interim we have selected predictor variables empirically, on the basis of significant association with logged equivalised household income at postcode sector. The test in regression models which follow are essentially of two types: a) how these associations change at different spatial scales, and how far they persist when household level characteristics are included in the model. The variables included are:

- % of working age economically active men and women who are unemployed
- % of households headed by a professional or managerial worker
- % of the working age population in poor health
- % of the population who have moved in last year
- Population density in persons per hectare

[4] There is an argument in the these of models for using a composite deprivation indicator. One of the reasons for not doing this here is that such indicators tend to be calculated for particular area sizes, and take into account the pattern of variation of the input variables at that area size. Recalculating such indicators for a flexible range of area sizes presents considerable problems.

One additional issue to be considered in the selection of indicators is the degree to which area and household level indicators should match each other. There are arguments for doing this, in part because it may become more straightforward to interpret the additional contribution of the local area component once personal characteristics are taken into account. However there are characteristics at both levels which may not naturally have a counterpart at the other level or where there may be difficulties in finding measures at both levels. Here there is some, but not complete correspondence between measures.

Results from area models are presented in tables 5, 6 and 7. in each table, the top panel shows models with area characteristics only, while the lower panel includes household characteristics. Table 5 shows results from standard geographies, ranging the very small scale Output are to the rather larger local authority district. Considering models with area variables only there is a clear decline in the degrees of model fit as we move away from small areas. However, even at district level three of the indicators (percentage unemployed, percentage professional and managers and in poor health) do remain significant. The one instance where an indicator is not significant at the smallest scale and becomes so for the postcode sector level is population density.

Once individual household characteristics are entered, associations with area characteristics tend to be attenuated, though at small scales several of the indicators remain significant. The percentage of households headed by a professional or managerial worker remains strongly significant at all spatial scales. It should be noted that this area indicator does not have a household level counterpart. Other exploratory work suggests that including some measure household socio-economic classification does reduce this somewhat, but does not remove the effect of the area component. At the larger scales this is the only area factor which remains significant.

Tables 6 and 7 use the bespoke areas. None of the spatial scales in these tables are as large as the two larger geographies used in table 5. We are looking at a somewhat narrower range of sizes, and hence the variations between the models are rather less. However, the pattern of declining model fit with increasing scale holds here too. It should be noted that the associations are stronger in the Output Area model in table 5 than in the smallest of the areas in table 6, which around twice the size the population size of an Output Area. In the top half of table 6 we again find the progressive decline in the significance of each area factor as the population size increases, with the exception of population density, which becomes significant at slightly larger areas. One can hypothesise that this is picking up some attribute of the wider environment. It should be noted that the coefficient changes sign as we move to larger areas. Once household characteristics are introduced it is only the percentage in the professional and managerial class and the percentage moving over the last year which remain significant in all models. The latter indicator is likely to be picking up inner urban areas of high population turnover. Once again population density is significant at large spatial scales.

The pattern is essentially similar when we define areas in terms of distance bands, shown in Table 7. The models small distance bands fit slightly more strongly for the smaller areas defined on distance rather than population size, but the difference is marginal. It would clearly be worth further investigation to clarify the relation between distance based units and population based units.

10 Conclusion

This paper has mainly been concerned to demonstrate the potential value of a combination of survey and spatial data, and within this context to present evidence on the spatial scale at which it would be most appropriate to undertake analyses of these data combinations. It involved two related but distinct pieces of analysis to illustrate this. Both focussed on household income in a new longitudinal survey as the 'dependent variable'. In the first case we looked at the role of the scale of spatial units in accounting for overall income inequality in a decomposition analysis. In the second case we examine models of the association between household income and local area characteristics at a range of scales.

The decomposition analysis showed clear evidence of an increasing proportion inequality accounted for by the between-group component of the decomposition as smaller scale areas were used. Moreover levels accounted for were comparable with that accounted for single dimensions of household characteristics. There are still issues to explore here around the implications of large numbers of spatial units used in the analysis, and it might also be worth examining more alternative decompostions, for example using classifications by area type which might overcome the issue of including large numbers of areas with small sample sizes.

In relation to the scale issues the evidence from the regression analysis of household income is also reasonably clear. Associations with area characteristics are stronger the smaller the area used, with the exception of population density, which seems most predictive at intermediate scales. However it should not be assumed that this will hold for all outcomes. Buck (2001) found different patterns of association with spatial scale for a range of different outcome variables. There is clearly scope here for more work, both to explore the associations with scale for different outcomes, and to investigate factors which may lead to these patterns of association.

Understanding of these factors would lead to a better basis for selecting variables to introduce into these models. It would clearly be possible to introduce different area characteristics at different scales, though this should be done with a clearer understanding of the rationale for their use.

There are also other potential lines of development of this work, using other measures apart from household income as defined here, including both alternative income concepts and for measures of poverty risk or other aspects of well-being.

Table 5: Regression models for log equivalised household income: standard geographies

	Output area		Postcode sector		Constituencies		District	
	Coef.	t	Coef.	t	Coef.	t	Coef.	t
Area variables only								
% unemployed	-0.01199	-8.26	-0.01134	-3.3	-0.01797	-3.69	-0.01803	-3.27
% professional and manager	0.01516	16.07	0.017133	10.18	0.015778	6.42	0.015726	5.99
% poor health	-0.01399	-8.33	-0.0158	-3.57	-0.01392	-2.39	-0.01248	-2.01
% moved in last year	-0.00254	-2.95	-0.00534	-3.17	8.98E-05	0.03	0.003309	0.82
Population density	0.003327	1.4	-0.01524	-2.38	-0.00533	-0.72	-0.00901	-1.22
Constant	7.242206	322.92	7.259532	163.73	7.234411	114.16	7.191672	104.6
R squared	0.0799		0.0469		0.0286		0.0245	
Area and household variables								
Social housing renter	-0.10785	-5.96	-0.17627	-10.21	-0.20104	-11.77	-0.20822	-12.2
Own with mortgage	0.223199	15.3	0.222206	15.16	0.224384	15.26	0.224648	15.29
N in employment in HH	0.35308	47.29	0.356723	47.7	0.357691	47.68	0.358694	47.82
N children	-0.12738	-17.76	-0.12626	-17.52	-0.12497	-17.29	-0.12555	-17.37
N aged 65 & over	0.065649	5.84	0.066746	5.93	0.070939	6.32	0.073373	6.54
Lone parent	-0.04324	-1.75	-0.05411	-2.18	-0.06244	-2.51	-0.06354	-2.55
% unemployed	-0.00298	-2.31	0.004525	1.53	0.000692	0.17	0.002755	0.59
% professional and manager	0.012638	15.41	0.016268	11.33	0.015513	7.43	0.016965	7.62
% poor health	-0.00402	-2.72	-0.00811	-2.14	-0.00919	-1.85	-0.00836	-1.59
% moved in last year	-0.00122	-1.63	-0.00435	-3	0.00038	0.13	0.001839	0.54
Population density	0.002738	1.33	-0.01304	-2.39	-0.00378	-0.6	-0.00608	-0.97
Constant	6.730513	278.42	6.72558	164.85	6.712263	119.97	6.664106	110.6
R squared	0.3122		0.3069		0.3011		0.3008	

Table 6: Regression models for log equivalised household income: Bespoke areas based on population sizes

	500 population		1,000 population		2,000 population		5,000 population		10,000 population	
	Coef.	t	Coef.	t	Coef.	t	Coef.	t	Coef.	t
Area variables only										
% unemployed	-0.01435	-7.11	-0.01616	-6.62	-0.01854	-6.4	-0.01927	-5.53	-0.02025	-5.13
% professional and manager	0.015632	13.45	0.016435	12.57	0.017099	11.86	0.018258	11.27	0.018243	10.3
% poor health	-0.01284	-5.31	-0.01318	-4.44	-0.01175	-3.35	-0.01065	-2.55	-0.01055	-2.31
% moved in last year	-0.00363	-3.28	-0.00515	-4.02	-0.0059	-4.06	-0.00766	-4.47	-0.00842	-4.37
Population density	-0.00026	-1.36	-5.5E-05	-0.26	5.14E-05	0.21	0.000412	1.49	0.000772	2.5
Constant	7.273285	253.14	7.281694	223.56	7.279293	200.05	7.267397	175.79	7.269461	160.14
R squared	0.0685		0.0668		0.0625		0.055		0.0484	
Area and household variables										
Social housing renter	-0.13041	-7.31	-0.13639	-7.7	-0.14779	-8.43	-0.16584	-9.58	-0.17793	-10.35
Own with mortgage	0.224729	15.39	0.223436	15.3	0.220966	15.12	0.219913	15.03	0.21998	15.02
N in employment in HH	0.354981	47.54	0.354144	47.41	0.354912	47.5	0.354857	47.46	0.355317	47.51
N children	-0.12707	-17.7	-0.12663	-17.63	-0.12633	-17.57	-0.12622	-17.53	-0.1261	-17.5
N aged 65 & over	0.061881	5.5	0.060647	5.38	0.061431	5.45	0.062274	5.53	0.063349	5.63
Lone parent	-0.04508	-1.82	-0.04787	-1.94	-0.04861	-1.96	-0.05274	-2.13	-0.05545	-2.24
% unemployed	-0.00102	-0.58	-0.00149	-0.7	-0.00216	-0.86	-0.0013	-0.43	-0.00196	-0.58
% professional and manager	0.01471	14.69	0.01558	13.86	0.016522	13.35	0.017416	12.57	0.017447	11.54
% poor health	-0.00314	-1.5	-0.0036	-1.41	-0.0027	-0.9	-0.00323	-0.9	-0.00378	-0.97
% moved in last year	-0.00258	-2.68	-0.00367	-3.31	-0.00445	-3.54	-0.0059	-3.99	-0.00728	-4.4
Population density	-0.00029	-1.76	-0.00013	-0.72	2.41E-06	0.01	0.000256	1.08	0.000667	2.53
Constant	6.727596	233.71	6.730272	211.84	6.720923	192.84	6.720553	173.82	6.732048	161.11
R squared	0.311		0.3108		0.3101		0.3089		0.3077	

Table 7: Regression models for log equivalised household income: Bespoke areas based on distance bands

	Within 250 metres		Within 500 metres		Within 1 Km		Within 2 Km	
	Coef.	t	Coef.	t	Coef.	t	Coef.	t
Area variables only								
% unemployed	-0.01434	-6.67	-0.01697	-6	-0.01766	-5.12	-0.01979	-5.36
% professional and manager	0.016966	13.55	0.016842	11.97	0.017541	11.07	0.017963	10.73
% poor health	-0.0124	-4.83	-0.0147	-4.39	-0.01398	-3.55	-0.011	-2.63
% moved in last year	-0.00397	-3.49	-0.00496	-3.77	-0.00646	-3.93	-0.00738	-4.01
Population density	-0.00036	-1.37	0.000238	0.78	0.000789	2.46	0.000811	2.61
Constant	7.26274	229.53	7.275265	203	7.262663	180.37	7.258304	169.35
R squared	0.0724		0.0644		0.0535		0.0496	
Area and household variables								
Social housing renter	-0.11232	-6.06	-0.13692	-7.72	-0.16629	-9.6	-0.17419	-10.11
Own with mortgage	0.223112	14.22	0.226991	15.09	0.221799	15.03	0.221827	15.12
N in employment in HH	0.360592	45.13	0.358323	46.62	0.357622	47.35	0.355901	47.49
N children	-0.13008	-17.08	-0.12838	-17.45	-0.12614	-17.41	-0.12616	-17.5
N aged 65 & over	0.061135	5.01	0.063374	5.43	0.065198	5.73	0.063962	5.67
Lone parent	-0.03783	-1.47	-0.04515	-1.8	-0.05088	-2.05	-0.05395	-2.18
% unemployed	-0.00177	-0.94	-0.0013	-0.53	-0.00073	-0.25	-0.00283	-0.89
% professional and manager	0.015347	14.26	0.015637	13	0.016586	12.27	0.017098	11.97
% poor health	-0.00253	-1.14	-0.00603	-2.1	-0.00589	-1.75	-0.00341	-0.95
% moved in last year	-0.00231	-2.36	-0.00371	-3.28	-0.005	-3.55	-0.00603	-3.82
Population density	-0.00044	-1.97	0.000184	0.71	0.000652	2.38	0.000732	2.76
Constant	6.716196	212.36	6.725622	195.17	6.717043	178.26	6.718154	169.23
R squared	0.3188		0.3171		0.3124		0.3087	

References

Berthoud, R. (2008) Area variations in household income across Britain, *Cambridge Journal of Regions, Economy and Society*, edition: 1, vol:1, 37 - 39, 2008

Bolster A, Burgess S, Johnston R, Jones K, Propper C, Sarker R, (2007) 'Neighbourhoods, households and income dynamics: a semi-parametric investigation of neighbourhood effects, *Journal of Economic Geography*, 2007,7, Pages:1-38,

Buck, N. (2001) 'Identifying Neighbourhood Effects on Social Exclusion' *Urban Studies*, 38, 2251-2275.

Cowell, F (1995) *Measuring Inequality*. 2d ed. Prentice Hall/Harvester–Wheatsheaf: Hemel Hempstead.

Dietz, R. D. (2002) 'The Estimation of Neighborhood Effects in the Social Sciences: An Interdisciplinary Approach' *Social Science Research*, 31, 539-575.

Durlauf, S. (2004) 'Neighborhood Effects' , *Handbook of Regional and Urban Economics*, Volume 4, J.V. Henderson and J.-F. Thisse, eds.,

Knies, G., Burgess, S., Propper, C (2008) 'Keeping Up With The Schmidts: An Empirical Test of Relative Deprivation Theory in the Neighbourhood Context', *Journal of Applied Social Science Studies*, 2008, volume 1

Manski, C., (1993) 'Identification of Endogenous Social Effects: The Reflection Problem', *Review of Economic Studies*, 60, 3, 531-542.

Manski, C., (2000) 'Economics Analysis of Social Interactions', *Journal of Economic Perspectives*, 14, 3, 115-136.

Pfeffermann, D. (2002). 'Small area estimation- new developments and directions', *International Statistical Review* 70, 125-143.

Rao, J. N. K. (2003). Small Area Estimation. New York: John Wiley & Son

Mookherjee, D. and Shorrocks, A. (1982) `A decomposition analysis of the trend in UK income inequality', *Economic Journal*, 92, 886-902.

Shorrocks, A (2004) 'Inequality Decomposition by Population Subgroups' *Econometrica*, 52, 1369-385.