Further details on the two sets of slides available for download from, respectively, Martin Ralphs from ONS’s Beyond 2011 team and Chris Dibben from St Andrews University

- ONS Beyond 2011 team are exploring a range of options to consideration of 2021 census or alternative approach and are due to report in 2014 with fully costed out CBA evaluations of these options plus pros, cons, risks and changes needed. Option one (some form of traditional census approach) is well understood by ONS and so options 2 & 3 are the focus of much of the Beyond 2011 workload in near future
  - Various census options (eg traditional, rolling, long-short, headcount & survey)
  - Administrative data options (eg aggregate data, partial linkage, full linkage)
  - Survey option: large scale survey & address register
- Key principles which sit above the debate are what is required from the data by different users across various key dimensions: timeliness; geographical scale; value (ie costs and benefits); precision; methodological simplicity & transparency; breadth of variables.
- Questions around what the key needs and demands placed upon the census are for government and for the various different user groups
- Beyond 2011 would like to receive information/research/evidence that would help them to estimate a cost-benefit analysis of these various options. How will the academic community (and other user groups, eg local authorities) coordinate the response to this call for evidence (eg ESRC)?
- In thinking about the key census variables, and the potential non-census methods to get these variables, it is helpful to separate the core population data that is needed by government (eg age bands by sex by local authorities) from the attribute data of characteristics. It may be possible/necessary to use different approaches to get these two sets of variables at small area level/micro level. If a traditional census (of whatever form) is not carried out and there is reliance on administrative data then which of these attribute data are attainable from these administrative data? Which are not and how will these be gained/estimated, to what scales and with what degree of precision?
- What core set of variables can be gained at small area level from linked administrative records? And what are the biases in coverage within any such linked admin datasets?
- In terms of thinking about what small area data SAE methods require then reducing the number of covariates is not necessarily highly problematic if the reductions come from variables that are highly correlated with variables that can be acquired. Hence, we need to also think about which covariates spatially vary independently of one another.
- What is the scope for linking the administrative data to the census spine?
- From ONS perspective there is a need for the SAE methods to be able to handle distributional issues as well as point estimates, to give a sense of likely error around estimates, and to estimate cross-classified outcomes, all of which the methods can cope with (though often do not)
- If admin data becomes more central then technical issue of how to ensure data security during the linkage process? (cf ADLS powerpoint) There are various models of remote and/or firewalled linkage to be explored. This raises a host of legal and ethical issues that may well need to be addressed if greater use of administrative and/or commercial data is desired. The report of the Administrative Data Taskforce will inevitably talk to these issues.
• Currently ONS have no legal powers to affect the data collection of government departments, resulting in risks around potential changes to administrative data sources (eg changes to which variables are collected or to their quality) which are key to ONS’ data provision and/or modelling.

• If admin data becomes more central then does this change the purpose and function of that data beyond its ‘routine’ departmental purpose? How does that affect the control of these data in legal terms? Does this, for example, open up scope potentially for ONS to request changes and/or improvements to the admin datasets collected across government? This may be relevant both to collect desired covariate data directly or to collect/improve variables within the admin data which are to be used for admin data linkage

• Variety of administrative data sources and key questions are around coverage, bias and linkage. Despite known issues with incomplete coverage of the population within it is also true in some areas admin data sources are stronger than census data (eg capturing homeless populations in the health data). How do we know about the biases in the administrative data? There are people who are clearly missing from these datasets because they do not interact with these systems. And these biases are not missing at random but rather are socially and spatially patterned: how would a coverage survey relate to this issue? There are also ‘too many’ people on these administrative datasets compared to census counts (esp in London) for a variety of reasons that would need to be dealt with (eg non-residents registered here for tax purposes, migration within the UK that the admin data have not picked up, people who have died overseas, people who have moved overseas, potentially UK deaths (if cannot connect death certification to administrative data spine))? Can these mismatches be predicted accurately? Can they be confidently rolled into the future?

• Given the above, an administrative data approach would almost certainly require some form of coverage survey to test and adjust the administrative data in future to known benchmarks to ensure accuracy. How large does this coverage survey need to be? This will depend on the scale down to which robust estimates are needed (with local authority at a minimum seeming likely). How frequent does it need to be, and is there scope to reduce its frequency assuming that calibration of the administrative data improves over time? How would the coverage survey identify the socially and spatially unequal biases in the admin data? And if one build adjustments to the administrative data into the process based on, for example, calibration against 2011 census then how can one check and, if necessary, update these calibrations into the future if no future census exists against which to benchmark? How can we ensure that the coverage survey can be linked back to the administrative data spine?

• Should any such coverage survey be fully enumerated in certain selected small areas or spread across all small areas? What difference would the sampling approach make for the ability to calibrate the small area data eg if robust only down to local authority level how would one generate sub-local authority adjustments for the different types of small areas or would one apply the same adjustment uniformly within the local authority? This is essentially a question of how to downscale accurately given the spatial variation at sub-local authority level. If one fully enumerates some small areas then could one apply these adjustments to similar types of small areas in other parts of the country (eg similar OAC areas or IMD deciles for example)?

• How should one seek to calibrate, and to what scales (eg local authority? Or smaller scale eg MSOA or LSOA)? One process might be to essentially ignore the processes by which admin and benchmark (eg census) data for (small) areas come to be out of line and just to seek ways to accurately and reliably calibrate those numbers. Another process however might be to seek to understand and model these patterns of migration, death etc in microdata so as to seek to predict population and...
sub-national level (probably local authority level). Such a process would require detailed empirical information about such processes however and it is not clear that these data exist at present. This approach would also require a second phase of calibration to known benchmarks at some spatial scale (local authority, or small area level) and so may introduce additional complexity for uncertain benefits. As noted above, the frequency of any coverage surveys may decrease over time as models learn how to calibrate accurately but there would need to be some way to ensure that adjustments remain accurate in the longer-term.

- In the absence of a unique ID across administrative datasets the linking of the admin datasets will inevitably rest on a combination of deterministic/cleric matching and probabilistic matching methods. Key questions arise therefore as to how accurate deterministic/clerical matching would currently be based on existing datasets? Which variables are key to such deterministic/clerical matching and what changes are required to improve the accuracy of these key variables? What percentage of cases can we (and, with data changes, could we) potentially match in a clerical manner of which we are pretty sure of its accuracy. If probabilistic matching is required (possibly only as a follow-up for cases unmatched after an initial deterministic/clerical round of matching) then which variables are key to boosting the likely accuracy of these matches? Are they present and well completed in the key datasets? What additional variables and/or processes might be needed/beneficial in future to better link these datasets?

- What is the scope for accessing commercial covariate data? Which companies would be most viable? How to ensure sustainability of any such commercial data sources?

- In terms of SAE the potential linkages are around:
  - The need for accurate covariate data is central to all SAE methods. But methods differ in the type of covariate data they need – spatial microsimulation methods & ecological regression approaches can operate with aggregate covariate data at the target small area geography whilst some of the statistical approaches require individual level micro-data for the small areas. Hence, SAE methods may be differently affected by any changes in the availability of covariate data.
  - The potential role of SAE in estimating either population data and/or attribute data down to small area level and/or refining the estimation processes used to generate these data.

- These conversations flow beyond the academic SAE community and have potential serious implications for all individuals and organisations interested/reliant upon robust small area data (eg local government). In terms of an admin data plus coverage survey approach, for example, it is likely a non-spatially clustered coverage survey may be robust down to local authority level only, leaving questions over how the precision of sub-local authority (and particularly the very small area level) covariate data (whether population or attribute data) can be ensured.

- If we are to move towards small area data with a degree of uncertainty around them, what size of error around these small area data is considered acceptable by different user groups (incl the public, government and academic/research community)? How can we assess this?