Office for
National Statistics

16th Meeting of the GSS Methodology Advisory Committee

19 May 2009

Myddelton Street • London

Hosted by the Office for National Statistics

Agenda • Minutes • Papers

16th Meeting of the Government Statistical Service Methodology Advisory Committee

19th of May 2009
FM13 Hawthorn, Office for National Statistics,
Myddelton Street, London, EC1R 1UW

List of contents

Office for National Statistics

1. GSS MAC 16 – agenda

Item	Time	Subject
1.	10.00	**Administration and information papers** GSS MAC 15: Minutes News from the GSS GSS MAC 15: Progress - summary and discussion
2.	10.30	**Paper 1:** Cost-benefit analysis of proposed new data requirements Authors: Craig B Orchard, Bronwen Coyle, Jacqui Jones, Sarah Green, ONS Discussant: Martin Weale
3.	11.15	*tea/coffee*
4.	11.30	**Paper 2:** Tackling biases in the dual-system estimator Authors: Owen Abbott (ONS), James Brown (Institute Of Education, University of London) Discussant: Jelke Bethlehem
6.	12.15	**Paper 3:** When to benchmark short term surveys to annual Author: Martin Brand, ONS Discussant: Ken Wallis
5.	13.00	*lunch: LE1 Regents*
6.	14.00	**Paper 4:** Developing an apportionment method for financial variables based on returned and synthetic local unit turnover data Author: Salah Merad, ONS Discussant: Sandy Stewart
7.	14.45	*tea / coffee*
8.	15.00	**Paper 5:** Developing expertise in record linkage within ONS Methodology Directorate Authors: Dick Heasman, Briony Eckstein, Peter Youens, ONS Discussant: Harvey Goldstein
9.	15.45	confirm date of the next meeting terms of membership AOB
10.	16.00	*Close*

Office for
National Statistics

2. GSS MAC 15 – minutes

Committee members present

Martin Brand	ONS	Frank Nolan	ONS
Robert Crouchley	University of Lancaster	Chris Skinner	Southampton University
Harvey Goldstein	University of Bristol	Sandy Stewart	Scottish Government
Rachel Leeser	Greater London Authority	Kenneth Wallis	University of Warwick
Jil Matheson	ONS	Martin Weale	NIESR

Presenters

Simon Field	ONS	Gareth James	ONS
Ruth Fulton	ONS	Alan Smith	ONS
John Hodgson	HSE		

Others present

Joanne Clements	ONS	Paul Smith	ONS
Jane Longhurst	ONS	Markus Sova	ONS
Louisa Nolan	ONS (secretary)	Kevin Stone	DASA
Steven Rogers	ONS	John Wood	ONS

Apologies

Jelke Bethlehem	Statistics Netherlands	Peter Lynn	University of Essex
David Hand	Imperial College London	Stephen Penneck	ONS
Graham Jenkinson	ONS		

Introduction

Frank Nolan opened the meeting and made introductions. The new committee member, Professor Robert Crouchley from Lancaster University was welcomed. Apologies were received from absent committee members.

The minutes from the 14th NSMAC meeting were approved without change. The change of title from NS to GSS MAC was noted, and it was agreed to keep the numbering of the meetings in the same order.

Comments on progress from NSMAC 14

Ken Wallis, who acted as the discussant on NSMAC 14 Paper 3: a state space approach to extracting the signal from uncertain data, noted that the comments and suggestions made about this paper at NSMAC 14 were for ONS as well as for the Bank of England authors, especially the point about methodological revisions. Paul Smith responded that ONS is currently working on this, and making progress.

Action for secretary

2a	obtain response on NSMAC14 Paper 3 from Paul Smith for GSS MAC 16

Comments and news from GSS / ONS

Frank Nolan added the following news items from the Census and Social Methodology Division to those presented in the GSS MAC 15 booklet.

Census progress

Since the last meeting of the MAC, good progress has been made with Methodology work related to the 2011 Census of Population. There have been meetings of the UK Census Design and Methodology Advisory Committee on October 22 and April 24. The most recent meeting discussed items on the Census questionnaire development, the Internet Questionnaire, the Address register development, Edit and Imputation strategy, Quality Assurance strategy, Evaluation of statistical disclosure control for tabular output, and the dissemination update.

Work is progressing well for the dress rehearsal of systems in October 2009. The dress rehearsal areas have been chosen and the questionnaires finalised. The main contractor for Census systems has been chosen and has started work.

Social Surveys

We continue to make progress with research into the question of sexual identity. There has been significant qualitative work here. This question is now running in the ONS Omnibus survey as a trial.

Work is also progressing on the disability survey.

Analytical Methods

Work is progressing on measuring uncertainty in the Index of Multiple Deprivation, a review of models for external immigration, and a review of the quality of demographic estimates (paper on agenda). Work is also being done on income estimation with the emphasis on households below the income threshold by Local Authority.

Work on disclosure control standards for microdata has been completed.

Recognition

Some work projects have received recognition for excellence with Alan Smith winning the Bo Sundgen Award at the International Marketing and Output Database Conference in Finland (1 - 5 September). We also contributed to projects which won ONS Excellence Awards - disability survey, sexual identity, and CommuterView.

Comments from the committee

Jil Matheson told the committee that work is beginning on ‘Beyond 2011’, a project looking at population estimates and the need for a census after the 2011 Census has commenced. She noted that while it is important not to undermine Census 2011 by prematurely informing the public that that might be the last census, a more public debate, possibly via the Royal Statistical Society, is planned. Martin Weale stressed the value of a longitudinal study, and Jil concurred.

Sandy Stewart asked if there was any update on the UK Statistics Authority’s assessment process priorities. Jil replied that the last meeting on the subject had identified the first ten sets of statistics to be formally assessed for National Statistic status. The list will be on the Authority’s website, together with a series of monitoring reports.

Rachel Leeser enquired about the coordination of statistics release policies across the GSS. She pointed out that different departments have different policies on microdata release, and that, although in theory, archive data was supposed to be available, in practice, this was not always the case. Martin Weale added that, following recent concerns about data loss, policies have become even more heterogeneous than they were, and there is no sense of a single overall policy. Jil Matheson noted this, and agreed that consistency of approach and an absence of artificial barriers were required.

2.1 An update on the methodological aspects of implementing SIC(2007)

Authors	Gareth James	ONS
Presenter	Gareth James	ONS
Discussant	Chris Skinner	Southampton University

This paper is intended to update the committee on progress made in the investigation of the methodological issues associated with the implementation of the change from the old Standard Industrial Classification (SIC) to the new SIC(2007). It aims to stimulate discussion among the MAC members about the proposed methodological approach to producing historic and current estimates under the new classification.

Discussion

The discussant, Chris Skinner, noted that this is an old and global problem, and that he was impressed by the thoroughness of the approach. He commented on the ad hoc nature of the conversion matrix approach to converting historic economic series to the new classification, but appreciated the logic behind it. He had no major suggestions for changing the overall approach.

He said that the proposed approach to the problem is analogous to the treatment of non-response, but in this case, complete out-of-scope domains exist. He was concerned that the underlying assumptions be clearly understood and openly stated, and believes that public honesty about these is the best policy. He was of the view that conversion at the lowest level (i.e. before aggregation) would be best.

Chris then went on to suggest that a cross-validation be performed. Half the data could be used to construct conversion matrices, and these could then be applied to the other half of the data and then compared with domain estimates. The data could perhaps be split by time.

At present, about 40% of the businesses on the ONS Inter-Departmental Business Register (IDBR) have had their SIC(2007) classification codes assigned by imputation, rather than via self-assessment. Chris suggested that it might be more reliable to use only the data from businesses with non-imputed codes to create conversion matrices.

The problem of variability in conversion matrices over time was addressed by a Canadian study, which found that four years of dual-coded data were required to get reliable matrices. As this is not currently available in the UK, one option might be to impute more dual-coding back in time for areas of special interest, e.g. mobile phones. An alternative method assuming a start time followed by linear growth for new industries does not sound very attractive.

Finally, Chris agreed that if the target and auxiliary variables are both turnover, they should be converted before deflation. He noted that the impact of this depends on how prices change in the areas of activity.

Comments and responses were then invited.

Martin Weale agreed with Chris that the more data are dual-coded, the more confident we are able to be about the stability of conversion matrices. He asked if it would be possible to top up the dual coding across several vintages of the IDBR. This would be done far enough apart in time to see seasonal patterns, for example, over three years. Gareth James replied that it was possible to impute the codes back in time from January 2008, but that this would be resource-intensive, and the resource was not available. Going forward in time, conversion matrices can be replaced with domain estimation when the system is in place. This should help with the estimation of levels, linking factors and discontinuities over the next three to twelve months. Martin was still concerned about seasonality in the conversion matrices, which would not be modelled by a single conversion matrix. Gareth agreed that it might be possible to create, for example, one winter and one summer conversion matrix, but that there would be issues with IDBR updates introducing inconsistencies.

Harvey Goldstein then asked if it was known how much uncertainty was introduced and propagated due to the imputation of coding of 40% of businesses on the IDBR. Gareth replied that there was not sufficient resource to investigate this. He noted that the 60% of businesses with non-imputed codes represent far more than 60% of total turnover, as it is the largest businesses for which text descriptions of economic activity exist.

Sandy Stewart expressed concerns over IDBR quality, and wondered how good self-assessment of classification is in practice. He wondered if it could be cross-checked with ProdCom. He also pointed out that the choice of reporting unit over local unit had a big impact on regional data, for example, if the a bank's headquarters moved south of the border, it would make a significant difference to Scottish statistics through the removal of the entire business from Scottish accounts. Finally, he took the view that conversion should happen using the best-quality raw data, aggregated to the highest level, and seasonal adjustment and chain-linking should be carried out at the end of the process.

Paul Smith pointed out that 60% of the UK economy is in the service sector, for which there is no equivalent to ProdCom, so checking the classification as suggested would not be possible across the whole economy.

Gareth thanked the committee for their responses. He agreed that dealing with new industry classifications was tough, and said that it might be necessary to go to economics experts for advice. He made the following responses to the committee's comments:

- businesses have a chance to correct their imputed assessment, and this happens naturally as part of the IDBR processes;
- cross-checking of classification with ProdCom is already done as a data-confrontation exercise, before annual updates are taken on to the IDBR;
- turnover for local units is not currently collected, so there is no opportunity to use local rather than reporting units at the moment. However, this situation is currently under investigation at ONS;
- he will undertake further investigation by splitting the historical micro-data into sub-samples and comparing them, if resources allow.

Suggestions to authors:

2.1a	cross-validation of the conversion matrices to be performed by taking sub-samples of the historical micro-data if resources allow

2.2 National Statistics and Web 2.0: new opportunities for turning statistics into knowledge?

Authors	Alan Smith	ONS
	Simon Field	ONS
Presenters	Alan Smith	ONS
	Simon Field	ONS
Discussant	Robert Crouchley	Lancaster University

This paper is intended to inform the committee of ONS' plans for using emergent internet technology and interactive visualisation to disseminate statistical information to a wider audience. The committee is invited to comment on these plans, and suggest further avenues of research.

Discussion

The discussant, Robert Crouchley, contributed the following points and questions.

In response to Question 1:

- if you miss Web 2.0, you miss out: this is a mega trend;
- it is the interdependencies between different sources of post-war data that he would find useful;
- do you know existing market demand? How is it used and by whom? Have you done a situation analysis recently?
- why stream instead of download?
- *who* will have increased 'understanding of life in the UK'?

In response to question 2:

- *who* will be empowered? For example, 'Joe the Plumber' – can he target his business? Can he download a regional economic forecast and disaggregate into local regions? He would need high level technical skills to do this (software and economics). It is more likely that he will overlay maps with descriptive statistics using someone else's simple application programming interface (API);
- in Web 2.0, people will make things up without evidence. Will there really be people who add new things? Web 2.0 is good at sharing existing knowledge, rather than creating really new information;
- there is a danger from loss of focus on contributions by experts, as this is overwhelmed by the general population's less-informed commentary.

In response to Question 3:

- an impact analysis is necessary, although these ignore self-selection and require a sophisticated model. It is difficult to assess how much impact use of Web 2.0 has had on 'improving the understanding of life in the UK'.

Comments were then invited from the committee

Martin Weale and Robert discussed how it was likely that data would be mis-attributed to ONS, in order to legitimise it to the community. This would lead to a lack of control and accuracy, with end users unable to identify what they were getting.

Harvey Goldstein agreed that all the negative issues that Robert had brought up were likely to occur. However, he wondered whether eventually (5-10 years?) a steady state would be achieved, and people would learn what to trust. In this case, would the end product would be so valuable, that the process would be worth it. He asked whether it was sensible to hope for this, and how it would be measured.

Robert said he did not believe a steady state was possible, because more information was entering the environment than we could ever process or understand. He pointed out that, although people said the same about the invention of the printing press, unlike printing, the web was available to everyone.

Harvey added that it was possible to maintain some control on the web. For example, some private Youtube groups exist. He also asked how these emergent technologies could be used in an educational setting, where it could be very valuable. Finally, he asked where the resources would come from to develop applications using the API, and whether it was ONS' responsibility to monitor this. It would need people with good technical skills, who may not be well-resourced and risk being drowned out by the wider, uninformed population.
Simon Field responded to the comments. He said that 'Joe the Plumber' may well not create applications using the API, but he would benefit from sites which have them. This happens already with e-bay.

He pointed out that the most stable web-based software available is Linux, which is open source software, written by a collection of collaborators. The software writers themselves may not be the main beneficiaries of their work. More study is required to know when a collection of collaborators / users reaches a critical mass for stability.

Simon thought that it ought to be part of ONS' mission to support the use of Web 2.0 to disseminate statistical data and information. He said ONS is used to working inside its own environment, but thought it ought to be a legitimate part of our work to enter into a wider debate, for example in BBC on-line discussions. Smart web solutions and data visualisations should be popularised for others to use. He noted that successful software was generally seeded by a single individual or organisation.

Martin Brand said he thinks that ONS has no choice. The mission statement says that we must promote the best use of our own data, but regulation must also be maintained. He asked whether there was a difference between sharing data and the use of data. For example, the Personal Inflation Calculator was developed by ONS staff. Users can enter rubbish into it, but it must be made clear that the data is not from ONS.

Ken Wallis commented that, in academia, work is monitored by peer review, and any abuse of data is pointed out in the publishing process. ONS, however, does not do this, and in fact often declines to act as referee on journal papers. ONS should perhaps consider how to rebuff the abuse of ONS data. Currently, abuse of data often goes uncorrected.

Harvey did not think it was feasible for ONS to correct, for example, the Daily Mail. However, he said that there was an opportunity for ONS to establish itself as a responsible authority. The best approach was to tell people where to go, and make use of the website easy when they get there.

Alan Smith said that it is increasingly unrealistic to assume that all of our external relationships with the media can be conducted via the press office in the existing fashion. Therefore, we really need to extend our thinking in this area. He added that he thought that educational use is one of the clinchers for embracing new technology, even allowing for Robert's objections. Work done with school children had shown that their attention span was longer and they could solve more complex problems using an interactive approach than they could using traditional methods. However, expectations should be managed. There is a lack of authoritative organisations involved in Web 2.0 development, as highlighted in the Gardener report.

Jil Matheson asked if it was possible to track data users, to which Simon Field replied that it was, and there were commercial applications to this. He noted that there were already sites which monitor, track and are able to ban direct users.

Sandy Stewart wondered whether something like the Personal Inflation Calculator would be useful as an auxiliary for regionalisation of data, if the obvious rubbish could be extracted. The consensus was that it would be a biased survey of the web population, algorithms for extracting the data would have to be based on what was already known, and it was unlikely to improve on existing regionalised data.

Suggestions to authors:

2.2a	carry out an impact analysis of current web activities

2.3 A simple method of computing a smooth non-linear fit to observations of known variance

Authors	John Hodgson	Health and Safety Executive
Presenter	John Hodgson	Health and Safety Executive
Discussant	Harvey Goldstein	University of Bristol

John Hodgson presented a method of smoothing independent Poisson variates that optimised smoothness with a constrained fit rather than the traditional approach of optimising fit with a trade-off with degree of smoothness. The constrained fit is based on a chi-squared statistic and the objective function for smoothness is based on squared second differences of fit. John also discussed possible variants of the fit constraint and the smoothness objective function and presented some results on the estimation of variance for the smoothed values.

Discussion

Harvey Goldstein offered the following points for discussion:

- The method is essentially non-parametric but he would prefer a parametric method, such as a regression spline. This would allow greater control over the smoothing procedure.
- The degree of smoothness depends on the choice of fitting constraint and it is not obvious what this should be.
- The validity of the process depends on the validity of the distributional assumptions made and the assumption that the 'true' underlying process is 'smooth'.
- The Poisson assumption is questionable because the data come from many workplaces.
- Other points were: the smoothed data do not look very smooth;
  how good is the algorithm used?
  how are the results to be used?
  has the method been compared with other, standard smoothing methods?

John Hodgson defended the Poisson assumption on the ground that the sum of two or more Poisson variates is also a Poisson variate. He also said that his method is easy to explain, applies a minimum of arbitrary decisions and the degree of smoothing depends on the data, not artificial constructs. He accepted that the choice of fitting constraint reintroduced an element of arbitrariness, but felt that the choice of median (or mean) overall chi-squared provided a "natural" solution.

Comment was then invited from other committee members.

Ken Wallis said that smoothing is similar to estimating trends for time series. This basically amounted to the long-established Henderson weighted average, with nothing much better appearing since.

Martin Brand asked about the effect of reporting errors. John Hodgson responded that the method responds to errors such as variation in reporting delays and the main problem relates to the underlying assumption of smoothness.

Chris Skinner questioned the use of fine stratification to justify the Poisson assumption because of the non-independence of simultaneous deaths. John Hodgson replied that the data should strictly relate to accidents, not deaths, though the number of multiple deaths fatalities in the time period shown was not enough to distort this assumption significantly. Chris then said that this method of producing a smooth fit while allowing for Poisson variation was quite neat.

Rachel Leeser said that the target statistic is a rate, which may be affected by changes in the denominator. Robert Crouchley made the similar point that the number at risk is changing (as the economy moves away from production to service industries), so the declining trend is to be expected. John Hodgson agreed but explained that the intention is to of the method was to obtain a high-level description, not to provide an explanation for the data.

Rachel Leeser asked what the effect would be of changing from financial year data to calendar year data. John Hodgson said that he had not examined this but wouldn't expect much difference, although there would be a need to reconcile the two different smoothed series. He also added that it would not be possible to smooth on a shorter periodicity than annual because seasonal differences would violate the smoothness assumption.

Sandy Stewart said that other methods, such as exponential smoothing, have the advantage of being able to change parameters to control the degree of smoothness. John Hodgson regarded this as a disadvantage because this control did not take account of the inherent variability of the data.

Harvey Goldstein said that any dependency between data points, such as that caused by variable reporting delays, would screw up the chi-squared fitting criterion. John Hodgson agreed that independence of data points was an essential assumption, but doubted that, overall, variation in reporting delays would introduce significant serial correlation.

Martin Brand concluded by saying that the method provides a useful tool for Poisson data more generally.

2.4 Measuring uncertainty in the Local Authority population estimates

Authors	Ruth Fulton	ONS
	Joanne Clements	ONS
Presenter	Ruth Fulton	ONS
Discussant	Rachel Leeser	Greater London Authority

This paper is an update on an ONS project established to improve the understanding, measurement and reporting of the accuracy of mid-year population estimates for Local Authorities. The paper outlines the overall approach that has been adopted. Particular issues that were addressed were: how quality issues were assessed; distributions of uncertainty estimated for each component of the population estimate; and how these were combined using simulation to provide overall indications of quality. A plan for further work is described, focusing on Internal Migration. Ruth Fulton requested feedback from the committee on the following issues:

- Overall approach, simulation methodology and composite quality measure
- Plans for further work including issues identified for Internal Migration and proposals for investigating these issues
- Existing sources of information, analysis or expertise on these issues

Rachel Leeser, the discussant, gave the following response.

There are no right answers to the questions posed by the authors.

It is important to address fundamental questions, such as why we want measures of uncertainty and how are they going to help users. Error estimates for national estimates are also important so that the Local Authority (LA) results can be set in context. The discussant recognised the complexity of the problem, but thought that quality measures for LA estimates by age/sex or for individual components of change would be more useful than just for LA totals. [Ruth Fulton responded that the intention is currently to investigate error measures for the total LA population and for important components of change. More detailed uncertainty measures could be investigated later depending upon the progress of this work.]

As the authors suggest, the error distributions are likely to be more complex than those initially tested. They are unlikely to be Normal (probably skewed) or proportional to the size of the component and may well be different for different components of change and for sub-populations (such as age/sex groups). All this could lead to different error distributions for different LAs.

Clues for assessing errors can be found in levels of error for past estimates or from unusual changes to current estimates e.g. in LA life expectancy figures, where sudden increases may be because migrant moves are being missed prior to death.

The simulation methodology used was appropriate, but correlations between components and systematic biases need to be considered. Errors for each component need to be addressed separately, with careful consideration of the sources of errors and the appropriate methods to combine them (e.g. multiplicative or additive).

If internal migration is focused on initially there will be some LAs (particularly in London) where this will not be very informative since international migration is the dominant component. By concentrating on the key issues, only part of the error on internal migration will be estimated. Although it is likely that these are the issues with the greatest impact, do they have the greatest uncertainty? Sensitivity analysis is required as validation of the error estimates is not possible.

Other points were: to consider the impact of different definitions of resident; interactions and correlations between international and internal migration; and the effect of Census low response on estimates of migration from Census.

The discussion was then opened up to the committee.

Martin Weale made the following points.

- He was pleased that ONS is addressing the question of reliability.
- He suggested using the Cauchy rather than the Uniform distribution, to avoid ruling out very large errors.
- 1,000 simulations are not nearly enough. Past experience suggests that 50,000-100,000 simulations are needed to obtain stability.
- Are there any constraints that can provide limits on the error distributions?
- If no other information is available, subjective impressions are better than nothing. Estimates should improve with practice and experience.

Paul Smith suggested that small area estimation methods might provide some guidance.

Harvey Goldstein said that it is essential to account for correlations. This is difficult but they need to be built into the simulations. Some experimental work could be carried out to identify the order of magnitude of the correlations.

Chris Skinner said that there are many uncertainties in this work, especially regarding correlations, but sensitivity analysis would help to identify the important issues, and should be a priority for future work.

Rachel Lesser suggested a detailed study of a particular LA where issues are known in order to inform the work further.

Martin Weale mentioned that combining evidence from different sources is similar to economic density forecasts where probability density functions are combined, by taking weighted averages. Ken Wallis responded that departures from normality found in some current density forecasts in macroeconomics are hard to pin down empirically; in non-normal cases the density of an aggregate variable cannot usually be obtained analytically from the densities of its component variables.

Martin Brand concluded by emphasising the need to consider the purpose of the work. Is it: to identify important errors; to identify areas for improvement; or to estimate errors for publication? The last goal is very challenging. Ruth Fulton responded that the target is the third option and agreed that although this is difficult, uncertainty measures could be summarised or banded.

Suggestions to authors

2.4a	consider alternative distributions for the error of different components
2.4.b	validate results against real examples where issues have already been identified
2.4c	investigate whether simulation should take into account correlation between errors of different components
2.4d	undertake sensitivity analysis within the simulation work

AOB

Terms of membership and committee member recruitment

It was agreed that a note should be circulated on the length of term of membership. Both Rachel Leeser and Martin Brand suggested that a diversity of experience, background etc should be considered when inviting new committee members.

Suggestions for future topics

Harvey Goldsmith suggested a paper on issues about pupil data bases linked across the whole education system, and in particular, how this can be used by government departments and external researchers. Frank Nolan said that some work was indeed already being done on this.

Martin Weale suggested a paper on changes to household surveys. Martin Brand put forward the idea of something on longitudinal weighting in the Longitudinal General Household Survey, or perhaps on wider issues of falling responses. Robert Crouchley agreed that missing data and non-response attrition would be interesting.

Action for secretary

2b	draft a note on Terms of Membership for the Chair to circulate

Summary of actions and suggestions:

Section	Participant	Action
2a	GSS MAC secretary	obtain response on NSMAC14 Paper 3 from Paul Smith for GSS MAC 16
2.1	Gareth James	cross-validation of the conversion matrices to be performed by taking sub-samples of the historical micro-data if resources allow
2.2	Alan Smith Simon Field	carry out an impact analysis of current web activities
2.4	Ruth Fulton Joanne Clements	consider alternative distributions for the error of different components validate results against real examples where issues have already be identified investigate whether simulation should take into account correlation between errors of different components undertake sensitivity analysis within the simulation work
2b	GSS MAC secretary	draft a note on Terms of Membership for the Chair to circulate

3. GSS MAC 15: progress

3.1 An update on the methodological aspects of implementing SIC(2007)

Progress on the methodological work required to change ONS business surveys and outputs to SIC(2007) has been good, with many achievements over the past few months. In particular, work has focussed on survey redesign and specification of methods and systems for domain estimation and parallel running. Responses from the GSS MAC have been helpful in guiding our approach. There is still a lot of work required in 2009 however, to enable the switch to take place smoothly, and to the planned timetable. This work has taken precedence recently, and due to limited resources, we have not yet been able to carry out cross validation of the conversion matrices by sub-sampling, as suggested by GSS MAC members at the last meeting. In addition, following discussion at a meeting of the SIC(2007) Implementation Project Board, it was confirmed that resources do not exist to allow dual-coding of units in the universe at any time prior to 2008.

Our report detailing recommendations for converting short-term statistics from one classification system to the other, which incorporates comments from the GSS MAC, has been accepted by Eurostat.

3.2 National Statistics and Web 2.0: new opportunities for turning statistics into knowledge

Further work on National Statistics and Web 2.0 is incorporated in the on-going iDissemination project.

3.3 A simple method of computing a smooth non-linear fit to observations of known variance

HSE has pursued work on smoothing along the lines discussed in a paper presented to MAC last November, widening the scope somewhat to look at a similar method developed in ONS methodology section and considering the suggestion from a MAC member to look again at exponential smoothing. We continue to believe that exploiting a priori knowledge about the variance of the observed observations to control the legitimate amount of smoothing is an important feature of the proposed method.

3.4 Measuring uncertainty in the Local Authority population estimates

Work has started on investigating key quality issues associated with the internal migration component of the population estimates (as documented in the paper we presented). The actions suggested will be taken into account in our analysis. We hope to publish the finding later on this year.

Progress on work done by the Methodology Directorate in response to the committee's comments on Paper 3, NSMAC 14, 'A state space approach to extracting the signal from uncertain data' will be presented as a verbal update at the meeting.

4. GSS MAC 16: news

Census

As reported to the last meeting the division is providing support to the Census in a number of key areas. The most important is on addressing, where an accurate Census Address Register is essential to the cost, efficiency and accuracy of the Census. It is even more important in 2011 as many returns will be posted out as well as posted back. Monitoring response and estimating for non response also depends on the reliability of the address register in identifying missing returns.

Over the last few months a pilot address register has been successfully developed and tested in selected local authority areas, demonstrating that the process works. Work has now started, using results from the pilot, on the national address matching process to develop the national register for use in the Census. This involves matching the two main address registers (Royal Mail's Postcode Address File and the Improvement and Development Agency's National Land and Property Gazetteer), taking account of other relevant information from Ordnance Survey and other sources, resolving anomalies with the address suppliers, performing on the ground checks in selected areas and then undertaking a final round of checks with local authorities.

Work has also been progressing on area planning, the printing of maps, developing the algorithms to define Census output areas and finally on providing material to enable the answers to Census questions on occupations and industry to be coded and processed.

GSS Statistical Policy and Standards Committee

This new Committee was established last September. Its initial work programme has focussed mainly on issues supporting the UK Statistics Authority's work especially on their Assessment process for National Statistics. Existing work on classifications, harmonisation and quality is being enhanced to ensure that standards are clear for use in the Assessment process. The UK Statistics Authority published their Code of Practice in January - they will be assessing National Statistics against this Code. Guidance notes are being prepared to support the Code, providing advice on action needed to ensure compliance. Finally GSS SPSC has been managing a Quality Improvement Fund, provided by the UK Statistics Authority, which the GSS can use to prepare for Assessment and do work on any recommendations emerging from Assessments.

Methodology Consultancy Service

Over the last year ONS has set up a new initiative - a Methodology Consultancy Service to provide expert methodological support, including training, to the rest of the GSS and the wider public sector. The service has made a sound start, with several significant projects completed, and is on target to earn enough income to cover its costs in 2009 / 10. Much of the Quality Improvement Fund money is being spent by other government departments with the service.

Geography services

ONS has completed the modernisation of the software systems underpinning its Geography services. This has improved access to geographical information to ONS users - it is now available direct from users' desktops. The internal geography production processes are also now based on the new system leading to more efficient and reliable product production.

ONS has been contributing to the implementation of the recently adopted "UK Location Strategy". We have a seat on the Location Council which is leading implementation. ONS work is important in ensuring that the increasing demand for geographical information and analysis is well supported.

GSS Methodology Conference 14

This year’s Government Statistical Service (GSS) Methodology Conference is being held on 30 June at Church House Conference Centre in London. The one day event is designed to bring together people working on methodological developments and applications from around the GSS, to share experiences and provide a forum for learning about different methods and techniques.

Confirmed speakers include; Karen Dunnell (National Statistician), Professor Sir Roger Jowell (Deputy Chair of the UK Statistics Authority), Richard Laux (UK Statistics Authority) and Graham Jenkinson (Methodology Directorate ONS). Parallel sessions are planned on 2011 Census, assessment and visualisation, classifications and indicators, using administrative data, statistics in finance, and imputation and simulation.

The conference is open to people working across the GSS and statisticians and other professionals who have a particular interest in official statistics. Details of registration and a provisional programme are available at the following page - http://www.ons.gov.uk/about/newsroom/events/fourteenth-gss-methodolgy-conference--30-june-2009/index.html

16th Meeting of the GSS Methodology Advisory Committee

Cost-Benefit Analysis of Proposed New Data Requirements

Craig B Orchard, Sarah Green, Bronwen Coyle, and Jacqui Jones

Office for National Statistics
craig.orchard@ons.gsi.gov.uk

Executive summary

The new Code of Practice (CoP) for Official Statistics specifies requirements for official statistics to follow and national statistics to adhere to. In the CoP there are two principles that point to the need for understanding the costs and benefits associated with new data requirements. This paper provides an overview of a proposal to meet this need via a cost benefit analysis model that would serve to provide an overall picture of the balance between the costs and benefits for economic surveys, which could facilitate informed decision making. It includes consideration of previous models, key components to include and dissemination of the information to meet the intended aim.

Aim of paper

- To provide an overview of a cost-benefit methodology for new survey data requirements and seek committee members' views and suggestions for improvement.

Requested actions from the committee

- General feedback on the cost-benefit methodology.
- Suggestions for improvement.

Main issues for discussion

QUESTION 1: Are these the correct components?

QUESTION 2: Are these the most appropriate measures?

QUESTION 3: How can we value the time taken to complete a questionnaire by households, individuals and communal establishments?

QUESTION 4. Is the use of discounting to obtain a net present value (NPV) for costs appropriate for inclusion in a CBA model to assess new data requirements?

QUESTION 5. Should we measure risk and if so how should it be integrated into the model?

QUESTION 6. How should we optimally select ‘other’ users so that the sample is representative but we are avoiding unnecessary burden?

QUESTION 7. Are we missing any salient benefits by limiting the model to consider only quality benefits?

QUESTION 8. Is the suggested equal weighting of quality dimensions in line with Eurostat guidelines an appropriate approach to take?

QUESTION 9. Should user and output manager views on quality be considered as being equal in value, or should greater weight be given to users since they ultimately define quality? If so how should the weighting work?

QUESTION 10. Should the CBA model developed for assessing new data requirements include an adjustment for the output priority?

QUESTION 11. If an adjustment for output priority is made, how do we balance between urgency and importance?

QUESTION 12. Should ONS produce a cash value for benefits or use the value produced for the change in quality alone to make an informed decision?

QUESTION 13. Based on the restrictions (inherent with CBA) of non-cash benefits/costs, do you think that the way we propose to summarise the outputs from the assessment tool for new data requirements is appropriate?

Office for
National Statistics

Cost Benefit Analysis of Proposed New Data Requirements

1. Introduction

With the launch of the new Code of Practice (CoP) for Official Statistics there are some new requirements for official statistics to follow and national Statistics to adhere to:

Principle 6, practice 4:
Analyse the costs of proposed new data requirements (to data suppliers) against the potential benefits.

Principle 7, practice 5:
Seek to balance quality (for example, accuracy and timeliness) against costs (including both costs to government and data suppliers), taking into account the expected uses of the statistics.

This paper provides an overview of a proposed cost-benefit methodology to assist with meeting these new requirements and provide a tool to give an overall picture of the balance between the costs and benefits which could facilitate informed decision making.

2. Background

The basic principle of Cost Benefit Analysis (CBA) is to weigh costs against benefits, hence the name. For CBA to be implemented successfully, it must show that a viable proposal is likely to be better than alternative possibilities, including the status quo. There are a number of CBA approaches and models and these have been used to inform the proposed CBA model. This section provides an overview of the other CBA approaches and models.

2.1. Approach 1

In 1965, Prest and Turvey defined the CBA process as ‘maximising the present value of all benefits less that of all costs, subject to specified constraints’. Their cost-benefit model was based on modern economic theory and outlined four key issues to address for successful CBA:

- Which costs and which benefits are to be included?
- How are the costs and benefits to be evaluated?
- Discounting of future benefits and costs over time to obtain a present day value
- What are the relevant constraints?

The drawback to this model is that it does not help to identify ways of measuring intangible benefits. Improvements to the accuracy or publication time of an output can be vitally important but it is likely that neither would result in any quantifiable financial gain. Benefits still, however, have to be evaluated.

2.2. Approach 2

In 1968, Marglin developed a model for CBA for use in a social context. This was broken down into two main components:

- A measurement of economic efficiency (those factors described in Prest and Turvey's model)
- A redistribution component (this is an important addition)

The redistribution component was designed to take into account different social significances. In effect, this is the application of unequal weights to the figures generated by the CBA so that the social importance of proposed changes can be adjusted as appropriate when weighing up the costs and benefits.

2.3. Approach 3

The Treasury has, for many years, provided guidance to other public sector bodies on how projects should be appraised, before significant funds are committed. This guidance is provided in the form of *The Green Book – Appraisal and Evaluation in Central Government* (HMT, 2003) and aims to ensure that no policy, programme, or project is adopted by public sector bodies before they have been effectively assessed against whether there are:

- better ways to achieve an objective
- better uses for these resources

As such *The Green Book* provides invaluable guidance into what is required for successful CBA.

The Green Book defines CBA as:

'analysis which quantifies in monetary terms as many of the costs and benefits of a proposal as feasible, including items for which the market does not provide a satisfactory measure of economic value'.

The Green Book sets out the process for appraisal and evaluation, focusing on the needs to:

- ensure that there is a clear identified need and that any proposed intervention is likely to be worth the cost. This overview must include an analysis of the negative consequences of intervention, as well as the results of not intervening, both of which must be outweighed to justify action (risk versus benefits)

- clearly identify the desired outcomes and objectives of an intervention in order to identify the full range of options that may be available to deliver them. Targets should be set to help progress towards meeting objectives
- carry out an appraisal of the possible options that may be available to deliver the desired outcomes and objectives. This is the CBA component
- use decision criteria and judgment to select the best option or options
- ensure that evaluation takes place to ensure that lessons are widely learned, communicated, and applied when assessing new proposals

2.4. Approach 4

The Cabinet Office's *The Magenta Book: Guidance Notes for Policy Evaluation and Analysis* focuses on the wider meaning of policy evaluation, as opposed to just economic evaluation, across government. *The Magenta Book* highlights a variety of analytical tools and methodological procedures from a wide range of academic disciplines. It defines policy evaluation as:

> 'a range of research methods to systematically investigate the effectiveness of policy interventions, implementation and processes, and to determine their merit, worth, or value in terms of improving the social and economic conditions of different stakeholders'

The Magenta Book stresses that the range of methods used is probably the most important factor for successful evaluation. It then goes on to discuss the different types of methods that can be used for policy evaluation (quantitative and qualitative methods, theory based approaches, research synthesis methods, and economic evaluation methods).

2.5. Approach 5

Following increased government focus on the costs of collecting statistical data over the last decade, the Bank of England (BoE) decided that the principles of CBA should be applied to its monetary and financial statistics. The result of this was the publication of the BoE guidelines in 2006 *Cost-benefit analysis of monetary and financial statistics: a practical guide* developed specifically for CBA of the BoE's statistics in relation to their statistical code of practice at the time.

The guidelines also include a CBA model for assessing requests for changes to existing statistics and requests for new statistical outputs and, importantly, introduces the concept that it is just as important to take into account the costs to respondents as it is to take into account its own costs. The model also includes methodology for assessing the benefits arising from any new data requests, although it does not go as far as placing an actual financial value on it.

Unfortunately, because the BoE CBA model was developed to cover requests for new statistical outputs as well as changes to existing ones, it is unable to provide the necessary level of resolution for informed decision-making since its remit is too wide. Furthermore, the BoE model makes the assumption that if an output is required by law, it must be more beneficial than one that isn't; although this is not always an accurate or reliable measure of benefits.

2.6. Current ONS Model

The 2006 ONS paper: 'A methodology for valuing statistical benefits' (Wallis, 2006) detailed the methodology for calculating non cash-releasing efficiency gains from improvements to ONS's key statistical outputs. The methodology developed provided a monitoring and assessment framework that ONS used to track progress towards the efficiency targets set following the Gershon Review. The method was endorsed by the Office for Government and Commerce (OGC).

The Wallis paper proposed a benefit-to-cost ratio methodology to value improvements to key statistical outputs. This methodology is based on calculating the baseline value of key statistical outputs and then applying a predefined ratio to this to obtain a monetary sum that 'represents' the value of improvements. The underlying principle is that the value of an improvement project to an output is a percentage of the output's total cost. Such a one-to-one relationship between costs and benefits could be seen to have limitations but, in terms of the use of this model, Wallis states that:

'clearly the assumption that cost equals value can be challenged, but for the purposes of calculating a baseline, using costs seems a sensible approach and avoids the need for the development of much more complicated techniques or extensive consultation with users of our key statistical outputs'.

To produce measures for the model, Wallis uses the six European Statistical Services (ESS) dimensions of quality, and a measure of risk, to give an improvement project a score against a dimension of quality, depending on the impact it is likely to have. The model's weights favour improvements to an output's relevance, accuracy, and risk; with the remaining dimensions given a lower weight.

The idea of aligning the potential benefits of an improvement project against the ESS dimensions of quality and calculating a baseline value of a statistical output when there are no apparent cash benefits are the key concepts introduced by Wallis in this model.

3. The Proposed Cost-Benefit Analysis Model for New Data Requirements

3.1. The Proposed Cost Benefit Analysis Model

To meet the CoP requirements, the proposed CBA model seeks to balance key elements of academic theory (Prest & Turvey, 1965; Marglin, 1968) with government guidelines on evaluation and appraisal (HM Treasury, 2003; Cabinet Office, 2003; Bank of England, 2006) and the previous ONS approach (Wallis, 2006). While the Wallis model was developed for ONS it was designed to monitor and assess efficiency savings, so it may be limited for a cost-benefit model developed to analyse new data requirements since user views are not included. We can build on the Wallis concept of aligning potential benefits of an improvement project against the ESS dimensions of quality, but to propose a CBA model to meet the requirements of the CoP will require the amalgamation of ideas from researched sources. This will ensure that an informed approach is taken; and that the framework developed addresses the more specific issue of assessing proposed changes to statistical surveys.

Returning to the requirements of the CoP, principle six, practice four clearly states:
Analyse the costs of new data requirements (to data suppliers) against the potential benefits.

While this recognises costs to suppliers it does not take into account the cost associated with collecting and producing statistics. As such, if it was only this practice we had to meet, we could identify benefits and know costs and go no further. However, under principle seven, practice five there is also a requirement to balance quality against costs to government and data suppliers:
Seek to balance quality (for example, accuracy and timeliness) against costs (including both costs to government and data suppliers), taking into account the expected uses of the statistics.

When taking both practices together the need for a cost-benefit analysis model to assess new data requirements becomes stronger and we can start to identify key components that the model will require. The components outlined in the Table 1 below are recommended for inclusion in the proposed CBA model for new survey data requirements.

Table 1. Recommended components for inclusion in the proposed CBA model

Need	• Is there a clear identified need?
Costs	• To respondents • To the survey organisation
Benefits	• Output prioritisation - How widely the output is used, who uses it, whether it contributes to policymaking, and whether the output is required by law • In relation to ESS dimension of quality
Weights	• Calculate weights to adjust the overall value of the calculated benefits and costs for the priority of outputs. • Adjustment of costs and benefits over time by discounting future benefits to obtain a present day value
Model limitations	• Identification of the relevant constraints of the model

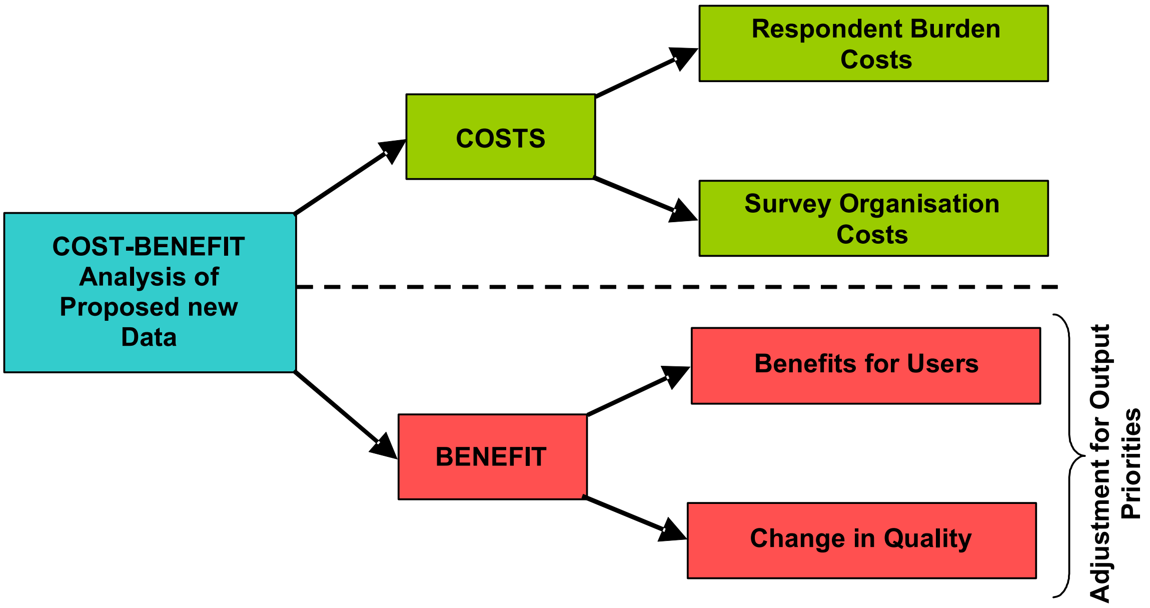
Further we can take from this that a model could be made up of four key measures:

- a measure of respondent burden costs
- a measure of survey organisational costs
- a measure of user benefits
- a measure of the change in quality

Drawing on the previous CBA approaches we are suggesting that an adjustment for the priority of outputs is made, based on the social importance redistribution weight used in the Marglin model.

The high level components of the proposed CBA model, and their interdependencies, are shown in Figure 1 below.

Figure 1. Key components of a CBA model for assessing proposed new data requirements


COST-BENEFIT Analysis of Proposed new Data
COSTS
Respondent Burden Costs
Survey Organisation Costs
BENEFIT
Benefits for Users
Change in Quality
Adjustment for Output Priorities

The proposed model assumes that all available options (e.g. availability of administrative data) will have been considered prior to a new survey data request being made.

QUESTION 1: Are these the correct components?

QUESTION 2: Are these the most appropriate measures?

3.2. Constraints to the Model

The suggestion in this paper is that the CBA model for new data requirement requests should specifically focus on those requests resulting in modifications to existing surveys. One of the limitations of some of the previous CBA approaches is the attempt to provide a 'one size fits all' model, which can reduce the relevance of the model's output. Constraining the model proposed in this paper to modifications, should allow for a more robust solution to these particular circumstances.

In addition there are a number of constraints that need to be considered as part of any CBA. For example, there would be no point in assessing a new data requirement request if there weren't the resource to implement it. Conversely, if the new data requirement request arises from a change in legislation, an organization may be required to make changes irrespective of the cost/benefit arising from it. Table 2 below indicates known constraints to CBA and establishes whether they are included in the CBA model this paper proposes.

Table 2. Constraints Associated with CBA and Inclusion in Proposed Model

Constraint	Brief Description	Included in Proposed ONS Model?
Strategic Impact	Potential improvement projects should be considered in terms of potential scale of impact, and how they fit with the organisation's general statistical output strategy.	Yes
Regulatory Impact	The impact on respondents of requesting information under new proposals should be assessed.	Yes
Legislation and Best Practice	Consideration should be given to pertinent legislation (both domestic and foreign). The impact of any relevant statutes (ie. the Data Protection or Freedom of Information Acts) should also be considered.	Yes
Information Management and Control	The information management and supporting IT that may be required for implementing a new data requirement.	No
Design Quality	The design quality of the modes of collection, questionnaire design, and overall process quality can be important in ensuring that objective(s) of an improvement project are successfully achieved.	No
Resource Constraints	Is an improvement project likely to meet resource constraints? This includes financial and time constraints.	No
Financial and Economic Rationale	Improvement projects need to be affordable to the organisation and need to be shown to be economically viable.	No
Partnering Arrangements	Proposals need to take account of how data are obtained. This includes third party sources collecting the information; whether they already collect similar data or not.	No

2.2. Measuring Respondent Burden Costs

Under the Prime Minister's instructions on the control of statistical surveys, government departments must minimise the burden that they place on respondents to business and local authority surveys. One of the measures taken to meet this obligation is to review business and local authority survey outputs on a triennial and quinquennial basis. During these reviews a measurement of respondent burden is undertaken by asking respondents a number of questions regarding the time taken to complete questionnaires, who completes it, and whether they require any external support. The collected data are then used to estimate respondent burden (cost to respondents). For requests to businesses and local authorities it is proposed that these data are used as a benchmark for survey managers to estimate whether implementing a new data requirement would increase the time taken to complete the survey questionnaire, change who completes it, and/or increase the time taken by an external book-keeper. The outputs from the survey manager's estimation would include an estimation of changes to overall respondent burden costs.

While the methods described above are established for measuring respondent burden costs to business and local authority there is no robust measurement for the respondent burden costs for surveys of households, individuals or communal establishments. The latter is a more difficult measure since while we know how long a response will take from a household etc. the question remains how to value the time. These issues will require further investigation before the proposed CBA could be extended further to non-business and non-local authority respondents.

QUESTION 3: How can we value the time taken to complete a questionnaire by households, individuals and communal establishments?

3.3. Measuring Survey Organisational Costs

In ONS, the cost of outputs is estimated as part of work planning and the ongoing commitment to meet efficiency savings, as set out in the Gershon Review. Current estimates of the costs associated with an output can be measured by using the full economic staff cost rates. These take into account the average annual rate of pay associated with each staff grade, as well as any associated overheads. As such, this ensures that operational costs, such as heating, electricity and equipment are taken into account in addition to costs associated with information management, national insurance, and pensions, known as shadow costs. This information can be used to estimate the running costs of a survey by accounting for time taken by the number of staff at specific grades to produce the existing output versus an estimate of the same variables to implement a new data requirement to the output.

In addition to the cost associated with the implementing the new data requirement, future costs should also be considered. For this to follow good accounting practice, the principle of discounting could be used to ensure that the project's running costs in future years are in today's money. Discounting takes the calculated costs for future years and applies a discount rate to obtain a Net Present Value (NPV) for future costs. Using an estimated NPV for a new data requirement in conjunction with an estimated implementation cost will result in an overall figure of the cost of a potential new data requirement. A potential new data requirement may seem more cost effective than an alternative, because it is cheaper to implement, but it may be more expensive in the long term due to higher running costs. More details on discounting and NPV can be found in *The Green Book* (2003) and Michel (2008).

QUESTION 4. Is the use of discounting to obtain a net present value (NPV) for costs appropriate for inclusion in a CBA model to assess new data requirements?

3.4. Measuring the Change in Quality

To measure a change in quality associated with a potential new data requirement, we need to have two measures:

- The current quality of the output
- The expected quality of the output if the new data requirement is added

Without knowing the current output quality, we would be unable to measure the change associated with a new data requirement.

Measuring statistical output quality is not easily quantifiable and is subjective since overall output quality, often viewed as an output's being 'fit for purpose', is ultimately defined by the user.

To have a consistent approach to measuring statistical output quality, ONS use the ESS dimensions of output quality given in the Table 3 below. Given that the proposed model aims to provide a tool for informed decision making, it is suggested that it may be of value to include a measure of risk with the dimensions. It would need to be considered how to integrate this into the model.

Table 3. Definitions of the ESS dimensions of quality

Dimension of Quality	ESS Definition
Relevance	A statistical product is relevant if it meets user needs.
Accuracy	Accuracy is the difference between the estimate and the true parameter value.
Timeliness & Punctuality	This is important for users since it is linked to an efficient use of the results.
Accessibility & Clarity	Accessibility is the ease with which users are able to access the data. It also relates to the format(s) in which the data are available and the availability of supporting information. Clarity refers to the quality and sufficiency of the metadata, illustrations and accompanying advice.
Comparability	The degree to which data can be compared over time and domain.
Coherence	The degree to which data that are derived from different sources or methods, but which refer to the same phenomenon, are similar.

It is proposed that relative measures of each dimension would be sought from both users and output managers in an attempt to provide a balanced view of overall quality. Output managers would give a considered view on each dimension of quality from a producer's point of view. Users would give their perceived view on quality which is vital to know if we are trying to produce statistical outputs that are fit for purpose. It is important to note that users will not just be confined to those requesting the new data requirement but will include the full range of potential interested users of the output such as academics, analysts and the press interested in assessing the effectiveness of government policy.

QUESTION 5. Should we measure risk and if so how should it be integrated into the model?

QUESTION 6. How should we optimally select 'other' users so that the sample is representative but we are avoiding unnecessary burden?

3.5. Measuring the Benefits to Users

The quality of statistical outputs is of great importance to users: this does not mean being of exceptionally high quality but being fit-for-purpose in the necessary areas. If a statistical organisation can change the quality of an output to meet user quality requirements better, then this will ultimately be of benefit to the users. Defining benefits to users as improvements in quality makes the relatively intangible concept of benefits measurable. This model limits benefits to users as improvements in quality.

When considering benefits to users the ESS dimensions (Table 3) can sometimes conflict. For instance, timeliness is in conflict with accuracy since accuracy generally takes time to achieve. If we are going to measure benefits for users and changes in quality via these dimensions it is important that we not only capture where there is an increase in one dimension but any negative impact this may have on the other dimensions.

QUESTION 7. Are we missing any salient benefits by limiting the model to consider only quality benefits?

3.6. Reconciling Views on Quality

Estimates of changes in quality and benefits to users would require weighting, since firstly all dimensions of quality need to be consolidated and secondly, output manager's and users' answers need to be reconciled.

In the Wallis (2006) CBA model, accuracy, relevance, and risk all carried greater weighting than other dimensions as it assumed that these were more important. Eurostat, however, recommends that each of the dimensions of quality should be considered equally (Eurostat, 2002). In line with Eurostat advice, it is suggested that equal weighting for all dimensions of quality, and risk also, is given in the proposed CBA model.

In terms of reconciling the output manager's and users' views on output quality, it is suggested that an average should be taken of the quality scores so that the overall value for current quality and the change in quality is a balance between the two groups. To ensure that user scores for quality are representative, the average scores for quality of those selected would be weighted equally against the output manager's scores.

QUESTION 8. Is the suggested equal weighting of quality dimensions in line with Eurostat guidelines an appropriate approach to take?

QUESTION 9. Should user and output manager views on quality be considered as being equal in value, or should greater weight be given to users since they ultimately define quality? If so how should the weighting work?

3.7. Adjusting for Output Priorities

If we are aiming to provide a consistent and comparable measurement of cost-benefit across outputs it is important that we in some way account for the relative priorities of outputs to one another. It is proposed to use the idea introduced by Marglin (1968) of adjusting for social importance.

Clearly any statistical outputs will have a priority order with each other, for example: an output on the economy that could influence government policy decisions will hold a higher priority in a statistical organisation than an output on a specific area of a small industry, that while important to the users does not have any wider impact. For statistical outputs, proxies for 'importance' could be how much an output is used, who uses it, and/or whether it is required by law or in policy making and such information about the new data request will be known by output managers. The BoE have designed a benefit assessment form which looks to assess the importance of outputs from a number of such proxies, from which this overall score for importance is achieved. It is suggested that we use a similar importance score for outputs. This would be applied as an adjustment to the consolidated measure of quality (benefits to users and changes in quality).

While we can apply Marglin's concept of social important to statistical outputs as described above, such an adjustment presents certain limitations to CBA in a statistical organisation. Assessing relative priorities is easiest when outputs are similar and inherently more difficult when they are different. For example, information on, say, retail sales or the balance of payments is important because it may need urgent action on fiscal or monetary policy. Information on population does not need immediate action but needs to inform longer term planning which is in some senses more important. Any adjustment must therefore take account of this balance between urgency and importance.

QUESTION 10. Should the CBA model developed for assessing new data requirements include an adjustment for the output priority?

QUESTION 11. If an adjustment for output priority is made, how do we balance between urgency and importance?

4. The Cost-Benefit Analysis Assessment Tool

Previous sections have outlined how the suggested components could be measured. The proposed tool to collect this information forms the user interface of the CBA model.

The tool comprises two main interfaces:

- an output manager questionnaire
- a user questionnaire

Table 4 shows the proposed content of each questionnaire, what it is designed to assess and who would be requested to provide the information.

Table 4. Contents of the CBA assessment tool questionnaires for output managers and users.

	Component to be Assessed				
Questionnaire Type	**Part A Current Output Quality**	**Part B Impact of the New Data Requirement**	**Part C Administrative Burden Costs**	**Part D Cost Survey Organisation**	**Part E Adjustment for Output Priorities**
Output Manager	x	x	x	x	x
User	x	x			

The questionnaires were designed in conjunction with the ONS Data Collection Methodology team and are currently being quality assured by a number of statistical output managers and users.

4.1. Collecting Information on Respondent Burden Costs and Survey Organisation Costs

It is proposed that estimates of respondent burden costs and survey organisation costs are collected from the output manager's questionnaire, parts C and D. These requests are restricted to the output managers only since, as discussed in section 2, information on these is readily available from management information and would not be known by users.

In estimating changes to respondent burden costs the proposed model bases calculations on the standard cost model method. The questionnaire requires output managers to enter the current respondent burden costs for their output and then assess potential changes in costs when new data requirements are included, based on:

- time to complete
- occupational category of respondents

Part D of the output manager questionnaire looks to collect the current running costs of the output and assess the level of extra resource required to include the new data requirements.

4.2. Collecting Information on the Benefits to Users and Changes in Quality

As discussed in section 2, it is proposed to measure changes in quality and benefits to users by assessing current output quality against the expected output quality, via the ESS dimensions, should the new data requirement be added. The sections on the current output quality and the impact of the new data requirement (parts A and B respectively), for both the output manager and the user questionnaires, provide the tool to collect the information needed to measure these.

Even though the same information is being sought from both output manager and users the content of the questions are tailored to each. An example of this can be seen in Table 5 below.

Table 5. Example differences between questions, based on the ESS dimensions of quality, between the output manager questionnaire and user questionnaire.

	Question	
ESS Dimension of Quality	**Output Manager Questionnaire**	**User Questionnaire**
Relevance	How often do you informally consult users to ensure that the information contained in the output meets their needs?	How often do you like to be informally consulted to ensure that the data are meeting your needs?
Accuracy & Reliability	What measures of accuracy and reliability accompany the outputs release when it is published?	What measures of accuracy and reliability currently accompany the output data you receive?

In addition, since it is unlikely that users could comment on whether a new data requirement would impact on the associated risks or timeliness and punctuality, these dimensions and aspects are not included in part B of the user questionnaire.

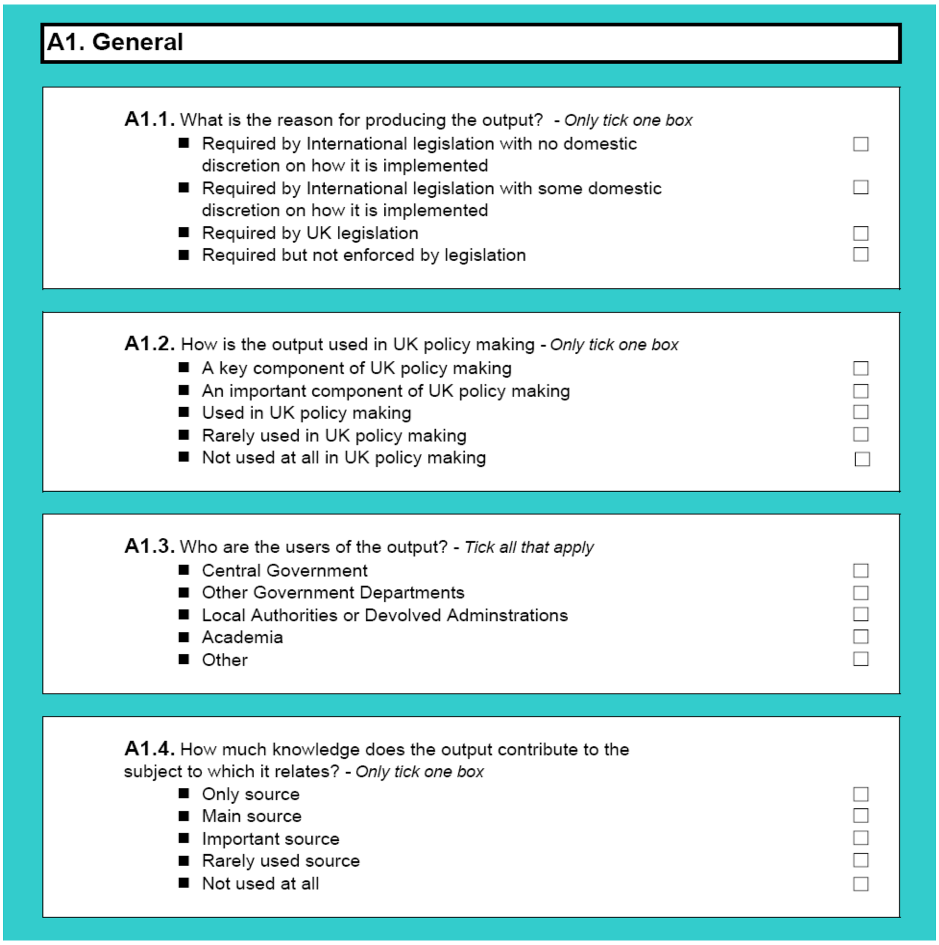
The questionnaires were developed to integrate the required weighting. The structure has been based on the dimensions of quality and within each section questions were carefully chosen to ensure that the information collected was not only useful but resulted in a balanced number in each. Moreover the scoring system to rate the quality questions (i.e. those questions in the sections on current output quality and the impact of the new data requirement; parts A and B respectively), are the same for both output managers and users. In this way, an effective scoring system for the current output quality and the change in output quality is proposed.

When designing the sections of the questionnaire to collect information on changes to quality, and hence benefit to users, a number of issues became apparent. For example, it was not possible to ask enough questions on coherence and comparability by themselves to provide enough resolution on these dimensions. Moreover, it was identified that any new data requirement for an existing output was more likely to result in a change in relevance or coherence/comparability. As such the coherence and comparability sections were combined together giving greater resolution in these sections.

4.3. Collecting Information to Adjust for Output Priorities

The BoE benefit assessment form (Bank of England, 2006), informed the development of proposed questions to assess the importance of outputs (see Figure 2 below). These questions are in the output manager's questionnaire.

Figure 2. Measuring the social importance of outputs in the output manager questionnaire. *Adapted from Bank of England (2006)*



A1. General

A1.1. What is the reason for producing the output? - *Only tick one box*

- Required by International legislation with no domestic discretion on how it is implemented □
- Required by International legislation with some domestic discretion on how it is implemented □
- Required by UK legislation □
- Required but not enforced by legislation □

A1.2. How is the output used in UK policy making - *Only tick one box*

- A key component of UK policy making □
- An important component of UK policy making □
- Used in UK policy making □
- Rarely used in UK policy making □
- Not used at all in UK policy making □

A1.3. Who are the users of the output? - *Tick all that apply*

- Central Government □
- Other Government Departments □
- Local Authorities or Devolved Adminstrations □
- Academia □
- Other □

A1.4. How much knowledge does the output contribute to the subject to which it relates? - *Only tick one box*

- Only source □
- Main source □
- Important source □
- Rarely used source □
- Not used at all □

4.4. Converting Non-Cash Benefits to a Cash Value

As with any model that doesn't measure cash benefits, it is extremely difficult to produce a monetary figure. The BoE CBA model (Bank of England, 2006) makes no attempt to apply a cash value to the statistical benefits it measures, but the HM Treasury guidelines (HM Treasury, 2003) recommend an 'informed' approach when presented with a situation where there are no cash benefits.

Currently, the only attempt at measuring statistical benefits in cash terms is the Wallis (2006) CBA model. This model was designed to give improvement projects a score against the dimensions of quality depending on its likely impact (see Table 1). This score was then used in conjunction with the total cost of an output to arrive at a cash value for benefits. Although this methodology is far from ideal, the same principles can be applied to the proposed CBA model for assessing new data requirements. The value obtained for the overall change in quality for implementing a new data requirement can be used with the total costs to obtain a value for the cash benefit.

QUESTION 12. Should ONS produce a cash value for benefits or use the value produced for the change in quality alone to make an informed decision?

4.5. Output from the Cost-Benefit Analysis Assessment Tool

The aim of the CBA tool proposed in this paper was never to give a single definitive figure on which to base decisions. The output from the proposed model is a summary of results that are hopefully informative, user friendly and easily interpreted. Figure 3 below is an example of the proposed summary sheet that could be produced from the tool once all the information has been collected and appropriately collated.

Figure 3. Example summary sheet produced from the CBA of a request for a new data requirement.

Cost-Benefit Analysis of a New Data Requirement Request for PRODCOM

Overview - Using a modified version of the existing cost-benefit analysis model used at ONS (ref), a cost-benefit model for the addition of questions to a survey has been developed. The implementation cost has been split over two years, the expected period that no further modifications will occur.

Implementation Cost	**- £ 100 k**
Running Cost (per annum)	**- £ 50 k**
Respondent Burden	**- £ 20 k**
Total Cost	**- £ 170 k**
Calculated Statistical Benefit	**- £ 400 k**

Overall Cost/Benefit -

£ 230 k

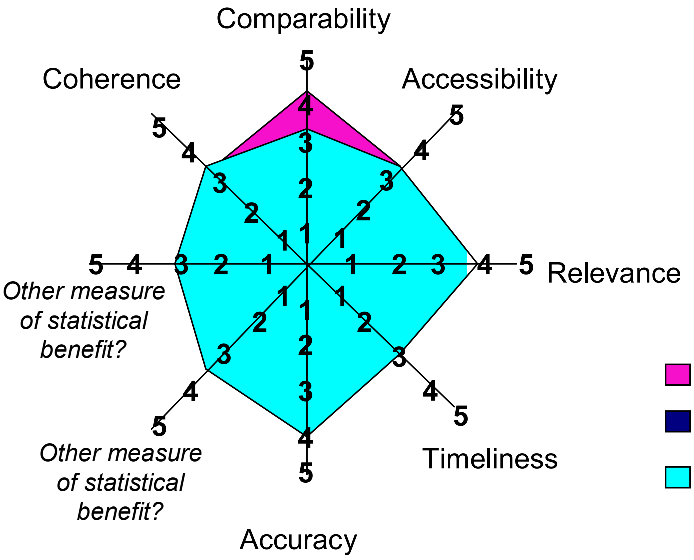
Adjusted Cost/Benefit* -

£ 150 k

*adjusted for social importance

Change in Overall Quality –

+ 5 %



KEY

- 🟪 Area of decreased quality
- ■ Areas where quality has remained static
- 🟦 Areas where quality has improved

- **Relevance**, the degree to which the statistical product meets user needs for both coverage and content, has **increased by 5 %**
- **Timeliness**, the lapse of time between publication and the period to which data refer, has **increased by 10%**
- Coherence, Accessibility, and Accuracy have remained static.
- **Comparability**, the degree to which data can be compared over time and domain, has been **reduced by 10%** since the additional questions have altered the way that x is collected.

QUESTION 13. Based on the restrictions (inherent with CBA) of non-cash benefits/costs, do you think that the way we propose to summarise the outputs from the assessment tool for new data requirements is appropriate?

5. Strengths and Weaknesses of the Proposed CBA Model & Tool

As with any model there are a number of strengths balanced by some limitations. These are considered in Tables 6 and 7.

Table 6. Strengths of the Proposed CBA Model & Tool

Strength	Consideration
Goes further than the Code of Practice requires	While this model and tool goes a little further than actually required, it does provide greater information about costs and benefits, relating one to the other, which could produce gains to a survey organisation's management information.
Consistent and comparable approach	The tool intends to provide a basis for informed decision making. In order to achieve this it is necessary to ensure that all statistical outputs are assessed on the same basis so that comparisons can be drawn when making decisions.
Uses experience of previous models	Previous models, and their interdependencies, have been extensively researched and as such the model proposed is based on their developments and looks to be further-reaching.
No single figure reported back	The model accounts for value being added not only by monetary savings but also by improvements in quality.
Indicates gains and losses in quality and costs	The model does limit the results to show only positive gains. In terms of quality this means that a trade-off between gains and losses can help inform a decision.
Questionnaire-based collection	The chosen collection method will collect consistent information from all respondents.
User-friendly interface	A single summary page highlighting main issues and values to consider make large volumes of information easy to interpret.

Table 7. Limitations of the Proposed CBA Model & Tool

Limitation	Consideration
Goes further than the Code of Practice requires	Could this be an unnecessary burden to survey organisations?
Subjectivity of Quality Measurements	Measuring quality is purely subjective and may be biased towards those selected to respond.
Figures may not be accurate estimates	The figures produced are not likely to be fully robust since the model was designed to provide a consistent approach that was comparable across outputs.
Assumes benefits will always be either cost savings or changes in quality	What if some users see the benefits of a new data requirement to a statistical output as further-reaching than quality or costs?
Questionnaire-based collection	Will increase burden on output managers and users.
Limited to analysis for business and local authority respondents	Due to the lack of robust measurement of respondent burden costs for households, individuals or communal establishments the model is currently only fully usable for analysis of new data requirements for business and local authority respondents.

6. Future plans

To date, the focus has been on developing an assessment tool that could evaluate requests for new data requirements within existing outputs, leading to informed decision making. One of the limitations noted in this paper is that the current model is sufficient for business and local authorities but a lack of robust measurement of respondent burden costs for households etc. means that it is not appropriate for analysing outputs that go to such respondent groups. Adpating the model to include these groups of surveys will be the next logical addition we want to make. Once this has been achieved, there are two further directions in which this work could be taken. Firstly, adapt the current model for new data requirements within existing outputs so that it can assess new data requirements that will require a new output. Secondly, adapt the current CBA model so that it could be used for any change that is carried out to an output, not just those arising from a new data requirement.

7. References

HM Treasury (2003). The Green Book. Appraisal and Evaluation in Central Government. The Stationary Office. *Her Majesty's Treasury*.

Bank of England (2006). Cost-benefit analysis of monetary and financial statistics, a practical guide. *The Bank of England*.

Cabinet Office (2003). The Magenta Book. Guidance Notes for Policy evaluation and analysis. *The Cabinet Office*.

Eurostat (2002). Quality in the european statistical system – the way forward. *Office for the Official Publications of the European Communities*.

Harberger AC (1978). On the use of distributional weights in social cost-benefit analysis. *The Journal of Political Economy*. **86:2**, p S87-S210.

Marglin SA (1968). The Discount Rate in Public Investment Evaluation *in* The Discount Rate in Public Investment Evaluation, Conference Proceedings, *Western Agricultural Economics Research Council, Denver.*

Michell RG (2008). Decision tools for budgetary analysis. *Government Finance Officers Association of the United States and Canada.*

Mishan EJ (1976). Cost-benefit analysis. *Praeger*.

ONS (2008). Simplification Plan 2008: Reducing the administrative burden caused by business surveys. *The Office for National Statistics.*

ONS (2009). Annual Report on Government Statistical Surveys. *The Office for National Statistics.*

Prest AR and Turvey R. (1965). Cost-benefit Analysis: A Survey. *The Economic Journal* 75(300):683-735.

UK Statistics Authority (2009). The Code of Practice for Official Statistics. *UK Statistics Authority*.

Wallis (2006). A methodology for valuing statistical benefits. *The Office for National Statistics*

16th Meeting of the GSS Methodology Advisory Committee

Tackling Biases in the Dual-System Estimator

Owen Abbott (ONS)
James Brown (Institute Of Education, University of London)

Executive summary

The 2011 Census will use Dual-System Estimation (DSE) as part of the methodology for assessing coverage, as it was in the 2001 Census. There are a number of assumptions that underpin DSE, and violation of those assumptions results in biased estimates of the population. This is often referred to as correlation bias, which is a key issue when using DSE. The experience in 2001 has shown that a strategy for estimating the levels of bias in the DSE is required, despite work to reduce the potential for bias in the first place. Previous work has recommended that the approach taken in 2001 should be extended together with further exploration of some of the potential sources of bias. Thus the strategy is to develop a global approach to estimating and adjusting for bias in the DSE alongside a number of studies of particular biases, such as the bias due to movers. These smaller studies will help to ensure plausibility of the global adjustments. In this paper we consider methodologies to both estimate and mitigate the sources of bias and hence outline the work that needs to be done to develop the methods.

Aim of paper

ONS is developing its strategy for tackling the issue of bias in the Dual-system estimator to be used in the 2011 Census methodology. The paper is presented for the opinion of the committee.

Requested actions from the committee

The committee is asked to provide any comments at the meeting, particularly to provide their views on the proposals in the paper.

Main issues for discussion

QUESTION 1: Does the committee have any other suggested approaches for QA of the closed population assumptions?

QUESTION 2: Does the committee have comments on the approach to estimating the bias in the DSE due to movers?

QUESTION 3: Does the committee have any ideas about how we could measure matching bias beyond the independent clerical matching proposed?

QUESTION 4: Does the committee agree that simulation of the more thorough treatment of the Chapman corrected DSE would be worthwhile?

QUESTION 5: Does the committee agree that matching the Census frame to the CCS listing is the most feasible approach for estimating residual bias at household level?

QUESTION 6: Is the committee aware of any similar work that might be relevant to extending the model for obtaining a person level odds ratio, or are there any alternative approaches?

QUESTION 7: The committee is invited to discuss reconciliation strategies and sources of data for checking the plausibility of the global bias adjustments.

Tackling Biases in the Dual-System Estimator

1. Introduction

Dual System Estimation (DSE) will be used in the coverage assessment and adjustment methodology for 2011, as it was in 2001. The high level methodology for 2011 is outlined in ONS (2008). There are a number of assumptions that underpin DSE, and violation of those assumptions results in biased estimates of the population (although it should be noted that DSE is not the only place where bias can occur, though others are beyond the scope of this paper). One of the critical assumptions is that the probabilities of being counted in the CCS are independent of the Census. If this assumption is violated it can result in a negative or positive bias.

In the 2001 One-number census a methodology for adjusting the dual-system estimates was developed and implemented, albeit late in the project. The strategy adopted was effectively a method for correcting all sources of bias in the DSE. The approach was to calibrate the DSEs to an external count at regional level. One of the key lessons was that bias is likely, and for 2011 a strategy must be considered and developed earlier.

Several papers have already considered the various potential sources of bias that can have an impact on the dual-system estimator. In addition, Abbott (2006) outlined the options for developing a strategy for assessing bias in the DSEs, recommending further development of the 2001 methodology, particularly around incorporating additional sources such as the census household frame. This was discussed at NSMAC (11). Additional work since NSMAC (11) has identified a number of potential studies of particular sources of bias, such as the bias due to movers in between the census and CCS. These are described in the paper.

Based on this previous work, our strategy is to develop a global approach to estimating and adjusting for bias in the DSE alongside a number of studies of particular biases, such as the bias due to movers. These smaller studies will help to ensure plausibility of the global adjustments.

In this paper we consider methodologies to both estimate and mitigate those sources and hence outline the work that needs to be done to develop the methods.

2. Background

This section provides some background on Dual-system estimation and the assumptions that underpin the methodology.

2.1 Dual-System Estimation

This paper assumes the reader is reasonably familiar with the application of Dual System Estimation, and therefore no derivations or theoretical background are given. For more comprehensive treatment see Brown (2000), Brown *et al* (2006) or Sekar and Deming (1949). A brief review follows, noting the assumptions underpinning the method.

Dual System Estimation is a standard method for estimating underenumeration. This was the approach used by the US Census Bureau following the 1980, 1990 and 2000 US Censuses, and the UK in the 2001 One-number census. Shortly after the census a post-enumeration survey is used to obtain an independent re-count of the population in a sample of areas. In the UK this is called the Census Coverage Survey (CCS). The application of dual-system estimation to a specific sub-group of a population combines these two counts to estimate the true population, allowing for people missed by both the census and the CCS, in the CCS sample areas.

The application of the DSE assumes that the following table can be created for the sub-population after matching the census and CCS data.

	Counted by CCS	Missed by CCS	
Counted by Census	$\mathbf{n_{11}}$	$\mathbf{n_{10}}$	$\mathbf{n_{1+}}$
Missed by Census	$\mathbf{n_{01}}$	n_{00}	n_{0+}
	$\mathbf{n_{+1}}$	n_{+0}	n_{++}

The counts in bold can be observed from matching the Census and CCS data while the counts in italics are a function of those individuals missed by both. The DSE of the unknown total n_{++} (which includes an estimate for the missed in both cell) is then given by $\hat{n}_{++}=\frac{n_{+1}\times n_{1+}}{n_{11}}$, or when the population counts are small $\hat{n}^{C}_{++}=\frac{(n_{+1}+1)\times(n_{1+}+1)}{n_{11}+1}-1$ by applying the Chapman correction (Chapman, 1951). The standard dual-system estimator is only asymptotically unbiased so when the actual population count n_{++} is small its bias can be (in relative terms) important. The Chapman correction removes the small sample bias from the DSE.

Re-arranging $\hat{n}^{C}_{++}$, and taking a first order approximation of $\left(1+\frac{1}{n_{11}}\right)^{-1}$ as $\left(1-\frac{1}{n_{11}}\right)$, we can express it as

$$\hat{n}^{C}_{++}\cong\hat{n}_{++}\left(1+\frac{1}{n_{+1}}+\frac{1}{n_{1+}}\right)\left(1-\frac{1}{n_{11}}\right)-\left(1-\frac{1}{n_{11}}\right)=\hat{n}_{++}\delta-\left(1-\frac{1}{n_{11}}\right)$$

(1)

provided n_{11} is greater than one. Details are in Appendix A. We will see later how this re-expression is important for the adjustment of the DSEs once the level of bias has been estimated.

The basic model behind the DSE is theoretically straightforward, assuming an underlying multinomial model for the four cells in the table of the matched data and independence between individuals in the population. All members of the target population must have a non-zero probability of being counted in any one of the four cells. Individuals with zero probability of falling into specific cells (and especially those with a zero probability of being counted in the margins) are not measured by the DSE, and in effect are out of scope, although they might be captured on other lists. In applying the DSE as an estimator of the unknown total there are two additional assumptions on the structure of the probabilities.

i) Homogeneity

The marginal probabilities of being counted by either the CCS or the census are homogeneous (or constant) across the target population (or domain). This is unlikely for most populations and results in bias.

ii) Independence

Unbiased estimation requires statistical independence at the individual level between the counting process of the census and the CCS. Brown *et al* (2006) provide a comprehensive explanation of this assumption and the impact of violation.

In addition to these assumptions regarding structure of the probabilities, the following two issues must be addressed if the model is to approximate reality.

iii) Accurate matching

It is necessary to match the two data sources to determine whether individuals on the two sources were counted once or twice. Errors in matching become biases in the dual system estimator (DSE).

iv) Closure

The census and the CCS must count the 'same' population. As they cannot physically count at the same time, both being in the field at the same time would have practical difficulties and likely violate independence, we have to deal with the issue of deaths and births between the Census and the CCS. For human populations, we can measure with reference to a specific point in the past, something that is not possible with animal populations, and gain a response with respect to the past from a member of the household present at both points in time. Therefore, although there will be births and deaths, they should not impact much as both counts refer to Census night, and the numbers of events will be relatively small. In addition, human populations migrate so care must be taken with movers between the two counting processes. Movers are more problematic as there may be no-one left at the second time point to record the existence of the individuals at the first time point.

Implicit in this assumption is that the two sources should be clean, with no individual being counted twice, or deceased individuals remaining on the register. If the sources are not totally clean (i.e. there is overcount in the Census or CCS), the population supposedly missed would be artificially high and the unknown population would be over-estimated.

2.2 Bias in Dual-System Estimates

Violation of any of the assumptions discussed above can lead to bias in DSEs. The resulting bias for each can be either positive or negative. This section explains how these biases occur, the likely magnitude and direction.

i) Homogeneity

The bias due to violation of this assumption is a function of the covariance of the probability 'pairs' (CCS and Census probability of being counted) – so the variance of each and correlation between the probabilities have an influence (which is why it can also be called correlation bias). Alho *et al* (1993) show that if the census and CCS inclusion probabilities are highly correlated and the inclusion probabilities for both Census and CCS vary a lot within the stratum, this results in a negative bias in the DSE. For some not unrealistic scenarios, the bias could possibly be on the order of 1.5 per cent. When one of the sources does have homogenous probabilities, the bias is zero (since the covariance becomes zero). If $\rho_{\text{cen,ccs}}$ is the correlation between census and CCS probabilities, σ^2_{cen} is the variance of the census probabilities, σ^2_{ccs} is the variance of the CCS probabilities and $\overline{\text{p}}_{\text{cen,ccs}}$ is the average product of the census and CCS probabilities across individuals, the following equation for the bias in a DSE demonstrates this relationship:

$$\text{bias} = -\frac{\rho_{\text{cen,ccs}}\sigma_{\text{cen}}\sigma_{\text{ccs}}}{\overline{\text{p}}_{\text{cen,ccs}}}.$$

(2)

ii) Independence

If people not counted by the census are less likely to be counted by the CCS than if they had been counted by the census (i.e. they change their behaviour in the CCS according to how they behaved in the Census) then this creates a negative bias. However, if people not counted by the census are more likely to be counted by the CCS than if they had been not counted by the census (e.g. they were not in the census so are more likely to be in the CCS because perhaps they feel they ought to) then this creates a positive bias. However, this is complex to quantify. Brown et al (2006) presents some scenarios and the resulting bias, which for realistic scenarios could be between -0.5 and -2 per cent.

iii) Accurate matching

If matches are not made (false negatives) then this creates a positive bias. A 0.5 per cent false negative match rate translates approximately into a positive 0.5 per cent bias.

iv) Closure

Bias due to movers was not really explored in great depth as part of the ONC, although the US has three different methods for estimating the bias. However, some recent work estimated that if the CCS were four weeks after Census Day then bias due to movers may be in the region of -0.13 per cent. The issue of overcount operates similarly to matching with any overcount feeding directly into positive bias. Thus a one per cent overcount will result in a one per cent positive bias.

3. The Closed Population

This section outlines how possible biases in the DSE due to the closure assumption are mitigated and how we will study movers, which have been identified as a likely problem in 2011.

3.1 Enumerating residents on census night

We first need to clearly define the population that the census intends to enumerate. This will be all individuals that are resident within England and Wales (or the UK) on census night where residence is determined by an intention to stay in the UK for at least three months (this will be six months in Scotland). This differs from 2001 in the use of three months intention rather than six months. There is also the important practical difference that the CCS will be commencing six weeks after Census Day rather than 3.5 weeks in 2001.

In reality human populations are not closed at a local level over time because individuals are born, individuals die, and individuals change location. Births are not an issue provided the CCS enumerators are clear that individuals must be residents on Census Day. Births can be literal births but could also be new migrants with an intention to stay beyond three months. There is the difficult issue of those that change their intention between Census Day and the CCS enumeration. When identifying residents the CCS will need to stress that we are counting with respect to intention at the time of the Census. Deaths of individuals can be captured provided not all household members have died (although the CCS enumerator needs to be sensitive when collecting this information). Again deaths can be literal but can also apply to those that have left the country. At the household level this will be an issue for single person households (particularly those at older ages) as well as short-term migrants leaving the country.

We will consider the possibility of QA information on these issues from death registration and ONS estimates on short-term migrants as well as stressing CCS training around the identification of residents as per Census Day.

QUESTION 1: Does the committee have any other suggested approaches for QA of the closed population assumptions?

3.2 Internal Movers

Internal moves of households essentially create ‘deaths’ in the area where they were on Census Day and ‘births’ in the new area for the CCS. In 2001 we judged that with the intensive Census fieldwork and the short gap this would not be an issue as the out-movers would be a minor increase in CCS non-response and the Census would count them as well as those that did not move implying a small increase in variance but no bias. Using the analysis of movers undertaken by the US Census Bureau in Griffin (2000), which treats movers as a source of heterogeneity bias, leads to a bias given by

$$-\frac{Td(1-c)(1-m)}{(1+dcm)(d+1)}$$

(3)

where T is the total population being estimated, $d = \dfrac{\text{number of movers}}{\text{number of non-movers}}$,

$c = \dfrac{\text{census coverage for movers}}{\text{census coverage non-movers}}$, and $m = \dfrac{\text{CCS coverage for movers}}{\text{CCS coverage non-movers}}$.

Looking at (3) we can see that the assumption in 2001 essentially assumed that while m equalled zero (all the movers were missed by the CCS) c equalled one resulting in no bias. Assuming c equal to one was sensible given the short and intensive nature of the field activity for the Census in 2001. Even when c did not exactly equal one any bias would be small because the short gap between Census Day and CCS fieldwork would also make d close to zero.

In 2011 this becomes a less realistic approach. The Census will have a more spread-out fieldwork process so may well miss those moving in the weeks just after Census Day at a higher rate than those that do not move. Therefore, if the out-movers are treated as non-response in the CCS, which will also be higher because of the increased gap, it will result in bias. Brown *et al* (2008) examined the options for dealing with movers, concluding that we should collect information on in-movers in the CCS sample areas (when the entire Census Day household has left) but the strategy for utilising this data is yet to be defined. Therefore, we need to develop a strategy that uses in-movers in the CCS areas to measure census coverage of movers by matching back to their Census Day residence. This will allow us to adjust the DSE at some macro-level for the bias from the differential census coverage of movers.

In the population, let us define the variable $M^{(a)}_{i,jg,kh}$ for individual i (from age-sex group a) resident in small area j of larger area g at the time of the CCS and small area k of larger area h on Census Day such that it equals one if the Census counts the individual and zero if the Census misses them. If j = k this identifies whether the Census counted or missed a non-mover while j ≠ k identifies whether the Census counted an out-mover from area k in the Census to area j in the CCS.

Within the large area g, the CCS will select a sample of small areas S_g and then a sample of individuals S_{jg} within sampled small area j implying an estimation weight of $w_{ij} = w_j \times w_{i|j}$ for i in the sample. For those responding in the CCS we will observe $M^{(a)}_{i,jg,kh} = m^{(a)}_{i,jg,kh}$ for i in the CCS sample S_{jg}, by matching within the area for the non-movers and by matching across the Census data in all areas j ≠ k for the movers (using the in-mover return in the CCS area). Therefore, an estimate of all out-movers (with age-sex a) from area k will be given by

$$\hat{M}^{(a)}_{kh} = \sum_{j \neq k} w_j \sum_{i \in S_{jg}} w_{i|j} \left(m^{(a)}_{i,jg,kh} + (1 - m^{(a)}_{i,jg,kh}) \right) \qquad (4)$$

where (4) is effectively the CCS based weighted sum of all the movers found in any sampled small area of the CCS that were both counted or missed by the Census in small area k. Summing over all k small areas will give an estimate $\hat{M}^{(a)}_{h}$ for out-movers from area h. This works because the CCS can estimate the total number of out-movers using its in-mover data;

we simply need to locate them back to their Census locations and then aggregate. An estimate within small area k of those counted by the Census before moving will be given by

$$\hat{N}^{(a)}_{kh} = \sum_{j \neq k} w_j \sum_{i \in S_j} w_{i|j} m^{(a)}_{i,jg,kh}$$

(5)

where (5) is based on the in-movers in the CCS that match back to a Census return in their Census location. Summing over all k small areas will give an estimate $\hat{N}^{(a)}_{h}$ for out-movers from area h counted by the Census prior to them moving.

Using the same data, but for the non-movers, an estimate of all the non-movers within area k (k in the CCS sample) will be given by

$$\hat{Y}^{(a)}_{kh} = \sum_{i \in S_{kh}} w_{i|k} \left(m^{(a)}_{i,kh,kh} + (1 - m^{(a)}_{i,kh,kh}) \right)$$

(6)

where (6) is based on those the CCS found that had not moved between the Census and CCS (j = k). A total for area h is then given by the weighted sum of the sample areas

$$\hat{Y}^{(a)}_{h} = \sum_{k \in S_h} w_k \hat{Y}^{(a)}_{kh} .$$

(7)

Equivalently, we can estimate the total non-movers found by the Census within area k (k in the CCS sample) using

$$\hat{X}^{(a)}_{kh} = \sum_{i \in S_{kh}} w_{i|k} m^{(a)}_{i,kh,kh}$$

(8)

leading to an estimate for area h given by $\hat{X}^{(a)}_{h} = \sum_{k \in S_h} w_k \hat{X}^{(a)}_{kh}$.

The estimators (4), (5), (6) and (8), with the corresponding estimates of totals including (7), will all be under-estimates because of non-response in the CCS. However, given that CCS response within an area will not differ between in-movers and non-movers (controlling for age-sex group) the proportional biases will be approximately equal[1].

Returning now to the bias due to movers given by (3), an approximately unbiased estimate of d, the ratio of movers to non-movers, for those in age-sex group a from area h in the Census will be given by

$$\hat{d}^{(a)}_{h} = \hat{M}^{(a)}_{h} \Big/ \hat{Y}^{(a)}_{h}$$

(9)

while an approximately unbiased estimate for c, the ratio of the census coverage for mover and non-movers, will be given by

$$\hat{c}^{(a)}_{h} = \frac{\hat{N}^{(a)}_{h} \Big/ \hat{M}^{(a)}_{h}}{\hat{X}^{(a)}_{h} \Big/ \hat{Y}^{(a)}_{h}}$$

(10)

[1] It would be possible to improve the estimates with respect to CCS non-response by calibrating $\hat{X}^{(a)}_{h} + \hat{N}^{(a)}_{h}$, the estimated Census count of non-movers and movers for area h, to the observed Census count for area h.

From these it is possible to estimate the relative bias due to movers for age-sex group a in area h as

$$-\frac{\hat{d}_h^{(a)}(1-\hat{c}_h^{(a)})}{(\hat{d}_h^{(a)}+1)}$$

(11)

where we should be concerned if the estimated variance of $\hat{c}_h^{(a)}$ suggests that c ≠ 1. Given that we have $\tilde{T}_h^{(a)}$, a biased estimate of the total population from our standard estimation strategy treating out-movers as non-response in the CCS, by combining (3) and (11) an unbiased estimate of the true population total T will be given by

$$\hat{T}_h^{(a)} = \frac{1+\hat{d}_h^{(a)}}{1+\hat{d}_h^{(a)}\hat{c}_h^{(a)}}\tilde{T}_h^{(a)}$$

(12)

where there will be little or no adjustment if c is close to one, the 2001 assumption that coverage is not different between movers and non-movers, and/or d is close to zero, the 2001 assumption of few movers between the Census and the CCS due to the short time gap.

In the discussion so far, we have referred to a generic geography of small areas (j and k) within larger areas (g and h). Within the structure of the CCS the small areas (j and k) will effectively be at the level of individual postcodes so that $M_{i,jg,kh}^{(a)}$ will pick-up the out-movers within postcode k, found elsewhere by the CCS, that the CCS would be forced to miss if it sampled postcode k. The larger areas (g and h) are an aggregation of postcodes to a level where the parameters c and d can be estimated, possibly the estimation areas within the standard CCS estimation or a higher aggregation to Government Office Regions given that the number of movers actually identified by the CCS, even within an estimation area, will be small.

QUESTION 2: Does the committee have comments on this approach to estimating the bias in the DSE due to movers?

3.3 Communal Establishments

The Census population also includes residents of communal establishments (prisons, student halls, army barracks, hotels etc). The large ones are pre-identified and treated separately by the Census. They are excluded from the CCS on the basis that the CCS is not likely to obtain any better coverage than the census.

The accuracy of these populations will be assessed in QA using other external sources and ideally these sources should be more than just overall totals but also give some idea of the age-sex profile.

The smaller communal establishments cannot be excluded from the CCS as their locations are not necessarily known when drawing the CCS sample. There can also be definitional issues around what a household is and what is small communal establishment is and so it is better for the CCS to include both rather than make a decision on inclusion based on a potentially different classification to the census. With matching (and based on the Census definition) we can identify the sub-sample of CEs within the Census and the CCS that can be excluded from the main estimation of the household population.

This sub-sample can then provide coverage information for this sub-population at some macro-level.

The appropriate level needs to be investigated given the information we have from 2001 regarding the numbers of these small CEs that are likely to be in the CCS sample.

4. Perfect Matching

The DSE relies on the ability to accurately match the counts of individuals from the two systems. Any failure in matching leads directly to bias. Finding matches that are not real (false positives) leads to a negative bias as the cell count n_{11} is inflated by the incorrect matches. Conversely, missing real matches (false negatives) has the opposite effect on the cell count n_{11} leading to a positive bias.

This was an issue that had to be dealt with in 2001. It was argued that false positives are unlikely as in the majority of cases the real match exists (as coverage rates in both the census and CCS are high) and this should have a higher matching weight because of the tightly defined blocking variables. In addition, the matching methodology uses the hierarchical structure of the data (individuals within households) which provides a powerful set of matching variables. Therefore, the main risk is missing matches. In 2001 the system automatically matched households/individuals with very high weights but clerical matching was used to determine match status for smaller weights with an emphasis on looking for the missed matches. The clerical matching process was strictly controlled and involved a number of layers of quality assurance to ensure the matching was of the highest possible quality. To explore whether there were any matching errors, this clerical component was undertaken independently by two different matchers with a reconciliation step where there were disagreements to eliminate (as much as possible) matching error. This assumes that matching error was purely due to clerical error rather than systematic methodological errors, which are almost impossible to measure without a third source of information that provides the true match. The target for matching accuracy was an error level of below 0.1 per cent. In 2001 the double matching estimated that the false negative rate was around 0.13 per cent, but of course these identified false negatives were then corrected. It therefore seems plausible that the false negative match rate was well below 0.1 per cent.

For 2011 the matching methodology will be similar to that employed in 2001 to again ensure the highest possible quality during matching, including independent matching to identify and remove any differences due to clerical error.

QUESTION 3: Does the committee have any ideas about how we could measure matching bias beyond the independent clerical matching proposed?

5. Overcount

As the 2001 Census operated a traditional approach in the field, overcount was not considered a major issue. The measurement approach from the CCS confirmed this but subsequent analysis (including the LS matching) suggests it might have been around 0.5%

(still a tenth of the undercount). The problem with census overcount is that they inflate the n_{10} cell leading to a positive bias in the DSE. In 2011 it is reasonable to expect that with changes in both the Census fieldwork processes and the structure of the population (more individuals with multiple addresses) this level will increase (similar changes in Canada between 2001 and 2006 saw an increase from around 1% to 1.5%).

Abbott and Brown (2007) outlined the approach for estimating adjustments to the DSE for overcount in the 2011 Census. This was discussed at NSMAC (13), and resulted in follow up work on the adjustments, reported by Brown and Abbott (2008).

As this work is reported on elsewhere and is designed to modify the Census counts that are used within the DSE (by effectively weighting them downwards), this source of bias is being dealt with through this process. Any residual is likely to inflate the estimates, and therefore will contribute within the global bias adjustment (although it will play off against negative biases). On the CCS side, we are assuming that the use of well-trained interviewers will essentially eliminate the issue, and any duplication will be identified by the matching process.

6. Residual Dependence and Correlation Bias

Sections 3, 4 and 5 have outlined how we will mitigate against bias in the DSE from a number of the sources that were identified in section 2, although these were mainly assumptions iii) and iv) (Perfect matching and closure). However, it is unrealistic to assume that the DSE will therefore be unbiased, particularly given the experience in 2001. This section outlines the methodology for estimating a global adjustment for any residual biases in the DSE which could potentially be from any failure of assumptions i) to iv) but in reality should mainly be from i) and ii) (Homogeneity and Independence).

A failure of either assumption i) or ii) will result in an apparent dependence between the census and CCS counts within the two-way table the DSE is based on. This apparent dependence will lead directly to bias in the DSE. The issue of movers discussed in section 3.2 is a form of correlation bias due to a failure of assumption i), the CCS completely misses the out-movers and the Census potentially counts them differently to the non-movers. We cannot distinguish between failures of i) and ii) in terms of impact on the estimates but the way we protect against failure is different. The term correlation bias strictly speaking applies only to a failure of assumption i) but as both are inter-twined we are using it to cover bias coming from apparent dependence regardless of the exact source.

To approximate ii), independence at the individual level, the two data collection processes are kept independent of each other in the field. In 2011 this ‘independence’ will be further strengthened by the very different enumeration strategies being applied in the Census and the CCS. Often considered the more difficult issue is approximating assumption i), the homogeneous capture probability. Work on the DSE for 2011 by Brown and Tromans (2007) identified a post-stratification approach similar to 2001 based on small areas for the DSE partitioned by age and sex as relatively effective at approximating assumption i). Further work on this has been undertaken and will be published during 2009.

We know from the 2001 experience (as well as our DSE simulation work for 2011) that there will likely be some residual correlation bias in the DSE coming from the way households respond to the CCS (given their Census experience) as well as not perfectly creating homogeneity. In 2001 an additional adjustment factor was applied to the DSEs at the end of the process to adjust for residual dependence and correlation bias as outlined by Brown *et al* (2006). This was estimated by comparing the estimated number of households coming from the DSEs with an external source to get an implied odds ratio between Census and CCS at the household level. This was then, through a simple synthetic model, adjusted to an odds ratio between individuals leading to the adjustment factor to be applied to the DSEs. We now need to look at developing an approach to allow similar adjustments for the residual dependence and correlation bias in 2011. Abbott (2006) outlined options for dealing with residual bias in the DSE, concluding that we should further develop the 2001 approach to be more flexible and therefore able to focus the dependence adjustment if needed rather than spreading it across all groups in the population. This was discussed at NSMAC meeting 11. This section therefore examines how the 2001 approach can be improved.

6.1 Improving the 2001 Approach

The basis of the adjustments applied to the 2001 data was an odds ratio (at a reasonably high level of aggregation) capturing the dependence between the Census and CCS household counts. This works because the standard DSE assumes the odds ratio is one and therefore an estimate of the missing cell n_{00} is given by $\frac{n_{10} \times n_{01}}{n_{11}}$ so that the DSE can be written as $\hat{n}_{++} = n_{11} + n_{10} + n_{01} + \frac{n_{01} \times n_{10}}{n_{11}}$. In general, when the odds ratio between the census and CCS is not one but is estimated to be $\hat{\gamma}$ the DSE can be adjusted to take this in to account leading to $\hat{n}_{++} = n_{11} + n_{10} + n_{01} + \hat{\gamma} \times \frac{n_{01} \times n_{10}}{n_{11}}$. As dependence only affects the estimate of the missed in both cell we can see that if both the census and CCS have high coverage the bias will tend to be small, even when the odds ratio moves substantially from one. Odds ratios greater than one result in a negative bias when using the standard DSE.

In 2001 we applied adjustments assuming we had used standard DSE and validated that this worked well, even when using the Chapman correction, through simulations. However, we can look a little more carefully at how to make adjustments when using the DSE with the Chapman correction. From (1) we can see that any adjustment for dependence to the DSE will feed through into the Chapman estimator such that

$$\hat{n}^{C}_{++} = \left(n_{11} + n_{10} + n_{01} + \hat{\gamma} \times \frac{n_{01} \times n_{10}}{n_{11}} \right) \delta - \left(1 - \frac{1}{n_{11}} \right).$$

(13)

From (1) and (13) we can see that if the dependence adjustment is defined as a multiplicative factor applied to the original DSE it is approximately the case that the same correction factor can be applied to the DSE with the Chapman correction, as was done with the 2001 adjustment. However, a 'more thorough' approach would account for the final term in (13) when making the adjustment, slightly inflating the final adjusted estimate (as the final term would have been adjusted by the same factor).

The effectiveness of this additional adjustment compared to the more simple approach taken in 2001 needs to be assessed using a simulation study.

QUESTION 4: Does the committee agree that simulation of the more thorough treatment of the Chapman corrected DSE would be worthwhile?

The approach used in 2001 relied on an estimate of the odds ratio between the Census and the CCS, $\hat{\gamma}$ based upon additional sources of data which were assumed to be the 'truth' (and therefore more plausible than the DSEs assuming an odds ratio of 1). The existence of the census household frame in 2011 has real potential to improve this component of the 2001 approach. Brown and Abbott (2008b) outlined how the frame could be used to create a triple system estimator that would allow a direct estimate of the odds ratio between the census and the CCS. The idea was to treat the original frame as one list and the actual Census as a second (dependent) list that would include additional households added as part of the enumeration process. Households would primarily have been added as part of a 100 per cent address check in the field as well as by householders phoning when they do not receive a form. The CCS would have provided the third list of households. There were some practical aspects that needed to be developed regarding this approach in relation to addresses identified as out-of-scope (vacant, second residences). However, given the way the frame is now developing, in particular the loss of the complete address check, it makes it difficult to conceptualise the frame as an additional source and not one totally tied to the Census. (In other words, the Census process will not across the country add households to the population not already on the frame to create the concept of the additional list.)
We do not wish to reject this as an approach to estimating the level of dependence at this stage, but are cautious as to whether it will be possible.

A different approach to getting at the bias is to create a lower bound on the estimate of n_{00}, the households missed by both the Census and the CCS. The existence of the frame allows this because after the Census fieldwork has been completed, there should be identified on the frame households that the Census enumerators failed to get a response from but which they believed to be occupied usual residences. This count will not be perfect (and some matching to the CCS will be necessary to adjust for Census mis-classification) but it should after adjustment give a good indication of the households that the Census knows exist but did not respond. After matching to the CCS responding households the residual should be less than the DSE estimate for n_{00}. If not it suggests residual dependence between the two counts and/or correlation bias that can be estimated and lead to an adjustment similar to the 2001 approach.
We think this represents the most feasible approach for estimating residual bias that uses the census household frame, although it does depend on matching and the underlying quality of the frame.

The approach outlined above assumes address level matching is possible between non-responding Census households on the frame and responding CCS households. An alternative would be to simply replace the role of the aggregated postal address file (PAF) in the 2001 adjustments with an aggregated 2011 Census address frame. This makes sense as we would expect the frame to be an improved count of households relative to the postal address file. The overall quality comes from the creation using additional information to PAF as well as the proposed address checking focused in to the 30 per cent of areas identified to be most problematic with respect to the creation of the address frame. In addition, a further one per cent random check of coverage across the whole frame is planned and this can be used to further enhance control totals calculated from the address frame. Using the household frame with a coverage correction will improve the estimate of the odds ratio $\hat{\gamma}$, and therefore the quality of the final adjustments relative to 2001 but perhaps not get as much benefit as the matching approach outlined in the previous paragraph or the full triple-system approach to get $\hat{\gamma}$.

QUESTION 5: Does the committee agree that matching the Census frame to the CCS listing is the most feasible approach for estimating residual bias at household level?

Once we have an estimate of the dependence between households this needs to be adjusted for the impact between individuals through the application of a synthetic model. The model used in 2001 assumed all households were size one (clearly incorrect although simulations showed it worked well as an approximation). This simplification was necessary given the short time-frame in 2001 when the approach was designed and implemented at the end of the processing.

There is scope here to attempt extending the 2001 approach to allow for differing household sizes (say one and more than one) given advanced planning.

QUESTION 6: Is the committee aware of any similar work that might be relevant to extending the model for obtaining a person level odds ratio, or are there any alternative approaches?

6.2 Other Sources of Information

One of the weaknesses of the 2001 approach was it made overall adjustments for residual dependence and correlation bias between households that were cascaded down to all the individuals. However, it could not focus in the adjustment on specific sub-populations. The adjustment for movers outlined in section 2 does focus to some extent within age-sex groups and if the bias is detected it should also be detected by the dependence adjustment strategy. The overall adjustment as applied in 2001 should have corrected for any residual bias from movers that existed as a result of treating them as non-response in the CCS as it was applied without any adjustment for movers. Therefore, (11) gives a lower bound on the population estimate for age-sex group a within area h after any dependence adjustment as the adjustment in section 6.1, applied to the DSE without any adjustment for movers, should not only pick-up on the correlation bias due to movers but also soak-up other sources of correlation and dependence bias. This allows us to assess whether the dependence adjustment is being focused in on the correct age-sex groups by comparison of different sources of information.

Work is needed to develop a reconciliation strategy for the scenario when the dependence adjustment within a particular age-sex group appears to be too low when compared to the other studies (such as the movers estimation) that identify sources of bias.

Other external sources also give us a way of checking whether the global adjustment is plausible for certain specific groups and hence whether the overall strategy is appropriate. Some of these sources may then be used, if they are accepted as being closer to the truth, to make additional adjustments after the global bias adjustments.

- Birth registration data is a key check on the estimates of babies. Inconsistencies between our estimates and the registration data are potential evidence of dependence within households, an issue not estimated by the 2001 approach that just considers residual bias at the household level.
- School census data helps with children of particular ages and perhaps the parents of those children if combined with dependency ratios.
- Matching with the main ONS household survey gives a real opportunity to measure within household dependence as we would expect the main survey to mainly suffer from non-response at the whole household level but provide a very good count of the usual residents within a counted household. **The ideal would be to provide a within household adjustment for broad age-sex and geography that could be combined with the household level dependence coming from the extended 2001 approach.** This will need some development and does assume access to the ONS household survey and an intensive matching exercise.
- The approach often suggested in the literature works by adjusting sex ratios and assuming the female counts are correct. In planning for 2001 this approach was outlined as the contingency for dependence in DSE. The approach explicitly requires reliable sex ratios from an external source and in July 2002 it was rejected as an approach for correcting the 2001 counts due to concerns regarding the sex ratios from the mid-year estimates. **ONS Centre for Demography are continuing to work on understanding sex ratio patterns in mid-year estimates and in administrative data sources, including patient register data, with the aim of having a sex ratio that could be considered sufficiently reliable to help in shaping adjustments to the Census estimates. The improvements being made to mid-year estimates may also give greater confidence to the sex ratios shown in the mid-year estimates.**
- Other demographic measures may also be useful in providing evidence for adjustments, although they are more likely to provide evidence for the plausibility of the census results than provide a precise method of adjustment to ensure a sensible demographic structure within the population. These include levels and trends in: age/sex specific fertility rates and summary measures; age/sex specific mortality rates and summary measures; age/sex in-migration rates; and, cohort change over time across censuses. Another measure, related to sex ratios, would be dependency ratios. As with the sex ratio approach, the issue is coming-up with viable estimates of demographic structures that are not heavily dependent on census data.

QUESTION 7: The committee is invited to discuss reconciliation strategies and sources of data for checking the plausibility of the global bias adjustments.

7. Further work

Work on developing the strategy for adjusting the DSEs for bias will be continuing during 2009, with a view to finalising proposals in October 2009. This work will include completion of:

1) Development of the overcount matching strategy and adjustments for the DSE.
2) Development of a similar matching strategy for CCS in-movers.

3) Development of the 2001 dependence adjustment model to be more flexible and to incorporate information on within household dependence, and evaluate its performance in a series of simulation studies.
4) Explore the sources of information to focus in on a plausible strategy for estimating the level of implied dependence (caused by both dependence and correlation bias), including a strategy for reconciliation.

8. References

Abbott, O. (2006) Options for adjusting Dual System Estimators for bias. Paper presented to 11th Meeting of the National Statistics Methodology Advisory Committee. Available on request.

Abbott, O. and Brown, J. (2007) Overcoverage in the 2011 UK Census. Paper presented to 13th Meeting of the National Statistics Methodology Advisory Committee. Available at www.statistics.gov.uk/methods_quality/downloads/NSMAC13_Census_Overcoverage.pdf

Alho, J. M., Mulry, M. H., Wurdeman, K. and Kim, J. (1993) Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, **88**, 1130-1136.

Brown, J. J. (2000) Design of a census coverage survey and its use in the estimation and adjustment of census underenumeration. University of Southampton, unpublished PhD thesis.

Brown, J., Abbott, O., and Diamond I. (2006) Dependence in the one-number census project. *J. R. Statist. Soc*. A, **169**, 883-902.

Brown, J. and Abbott, O. (2008a) Response to Discussion at NSMAC on Estimation of Over-Count. Note for 14th Meeting of the National Statistics Methodology Advisory Committee. Available at www.statistics.gov.uk/methods_quality/downloads/Estimation_overcount_followup.doc

Brown, J. and Abbott, O. (2008b) Ideas for Estimation in the 2011 Census – Utilising the Census Address Frame. Paper presented to the UK Census Design and Methodology Advisory Committee (UKCDMAC). Available on request.

Brown, J., Abbott, O. and Taylor, A. (2008) Options for dealing with movers in between Census and CCS. Internal research paper. Available on request.

Brown, J.J. and Tromans, N. (2007) Methodological Options for Applying Dual System Estimation. Paper presented at ISI satellite conference. Available at www.s3ri.soton.ac.uk/isi2007/papers/Paper22.pdf

Chapman, D.G. (1951) Some Properties of the hypergeometric distribution with applications to zoological censuses. *Univ. Calif. Public. Stat.* **1**. 131-160.

Griffin, R. (2000). Accuracy and Coverage Evaluation: Dual System Estimation. DSSD Census 2000 Procedures and Operations Memorandum Series, Q-20, US Census Bureau.

ONS (2008) 2011 UK Coverage Assessment and Adjustment Methodology. Census Advisory Group paper AG (08)05. Available at www.ons.gov.uk/census/2011-census/consultations/user-adv-groups/census-adv-groups/statistical-dev/2011-uk-cen-cov-ass.pdf

Sekar, C. C. and Deming W. E. (1949) On a method of estimating birth and death rates and the extent of registration*. Journal of the American Statistical Association*, **44**, 101-115.

Appendix A

$$
\begin{aligned}
\hat{n}^{C}_{++} &= \frac{(n_{+1}+1)\times(n_{1+}+1)}{n_{11}+1}-1=\frac{n_{+1}\times n_{1+}+n_{+1}+n_{1+}+1-n_{11}-1}{n_{11}+1}=\frac{n_{+1}\times n_{1+}+n_{+1}+n_{1+}-n_{11}}{n_{11}\left(1+\dfrac{1}{n_{11}}\right)}\\
&\cong \frac{n_{+1}\times n_{1+}+n_{+1}+n_{1+}-n_{11}}{n_{11}}\left(1-\frac{1}{n_{11}}\right)=\frac{n_{+1}\times n_{1+}+n_{+1}+n_{1+}}{n_{11}}\left(1-\frac{1}{n_{11}}\right)-\frac{n_{11}}{n_{11}}\left(1-\frac{1}{n_{11}}\right)\\
&= \frac{n_{+1}\times n_{1+}}{n_{11}}\left(1+\frac{1}{n_{1+}}+\frac{1}{n_{+1}}\right)\times\left(1-\frac{1}{n_{11}}\right)-\left(1-\frac{1}{n_{11}}\right)
\end{aligned}
$$

16th Meeting of the GSS Methodology Advisory Committee

When to benchmark short term surveys to annual

Martin Brand, ONS

Aim of paper

This paper examines the issue of whether or not to benchmark sub-annual (usually monthly) surveys to annual. It concludes by suggesting a possible policy for ONS surveys.

Requested actions from the committee

GSSMAC is requested to discuss criteria for benchmarking business surveys and the policy ONS should adopt. There may be parallels with other fields of application when two estimates have to be reconciled.

Main issues for discussion

Question 1: Is the list of criteria for deciding when to benchmark correct?

Question 2: Is the list of pre-requisites for benchmarking correct?

Question 3: Does GSSMAC concur that the ONS benchmarking policy should not be global ie it should be selective, judging each case on its merits?

When to benchmark short term surveys to annual

1. Introduction

ONS conducts a range of sub-annual and annual surveys business. Sub-annual are usually designed to measure change whilst annual are principally designed to measure levels.

In some cases, but not all, the concepts and target populations coincide and there is an evident inconsistency between annual estimates based on sub-annual versus annual. This can cause problems for some users.

It is possible to align the two sources, usually by "benchmarking" the sub-annual to the annual. However, there is no stated ONS policy for when this should be done – survey results areas have evolved practices independently.

The purpose of this paper is to promote discussion on the *policy* of when to benchmark. Excluded is a discussion of the technical merits of different competing *methods* of benchmarking once the decision is made to benchmark.

2. Background

In terms of scope, what we are describing here is benchmarking one Sources survey with another. Excluded is the wider issue of consistency between Sources outputs and National Accounts series. As an example, this note covers the benchmarking or not of the Monthly Production Inquiry (MPI) to the Annual Business Survey Part 2 (ABI2), but not the relationship of the MPI to the Index of Production (IoP) or the IoP to annual constant price data within the National Accounts framework.

Some consideration of this issue took place up to 2000 within the NASSG (the Sources and Analysis liaison group for National Accounts and Surveys). At that time, the idea of benchmarking sub-annual surveys to annual appeared to be in favour, for example to ensure coherence.

Note that the purpose of short term surveys is usually the measurement of change; hence producing series without discontinuities is often very important. This can jar with annual estimates which are about the best picture for a year - which means that change between two annual points will include things other than "genuine growth", such as reclassifications.

2.1 Criteria for deciding when to benchmark

In principle, differences between two surveys may occur because of different approaches at each and every stage of the Statistical Value Chain. For example, from the fundamental concepts, the questionnaire and mode of collection, right through to final estimation.

Clearly we would expect two estimates from two separate sample surveys to be different because in general the sample selected (and response rates) will be different.

It may be useful to consider three types of differences: Conceptual, Methodological and Operational (denoted C, M and O below).

It is proposed in this note that there are a number of key considerations where, if the annual survey is thought to be significantly superior, that might support a case for benchmarking. The key criteria proposed are as follows:

C Inferior variable concept. It may be that the short term survey does not collect the precise concept required and that it is in a sense a "proxy" to a better concept measured in the annual survey.

C Inferior breakdowns (eg. variables or local units). The short term survey may be used to estimate a sub-annual path for more detailed variables, units or geographies measured by the annual survey.

M Inferior estimation (eg. matched pairs). In some cases, the estimation method used may be inferior. For example, it is well established that matched pairs runs a significant risk of drift in the medium term.

O Inferior coverage of the population. This occurs when the sub-annual survey has an inferior population coverage compared to the annual survey. Typically, this will be when certain parts of the IDBR are omitted, for example small firms or particular industries. Less commonly, it may be that the register used for the short term surveys is inferior in terms of coverage to that used for the annual.

O Inferior precision (including response rate). This is where because of smaller sample sizes (or lower response levels), the sub-annual survey has inferior precision to the annual.

O Inferior measurement at survey time. In some cases, although the variable definition is the same, it may not be possible for the respondent to accurately measure the variable within the short time scale available for sub-annual surveys. For example, while some financial variables can be estimated sub-annually, the annual figures - when company accounts are finalised - are considered to be much more accurate.

O Inferior non-sampling errors. This would occur when the sub-annual surveys have higher non-sampling errors eg. if editing was cruder, or less cross-checking applied.

User Consistency is paramount. In some situations, consistency between short term and annual surveys is considered paramount by users; inconsistency cannot be tolerated. If a short term survey is used for levels as well as change, then this may increase the case for benchmarking.

There are of course other reasons why sub-annual and annual surveys might be different. Examples are given below - it is suggested that these are less strong reasons to benchmark than the key reasons given above:

M Design eg. strata. There can be detailed design differences eg. in industry and employment stratification.

O Frame, timing, births, deaths. Annual surveys are selected later than monthly - at the end of the reference year rather than through the year. It is therefore possible that the IDBR classifications may be superior for annuals. On the other hand, sub-annuals may handle in-year births and deaths better - and, if the desired coverage is all firms who traded in the year, may be superior. However, it is suggested that these are not major factors in determining benchmarking.

O Mode of collection. Differences can occur due to the mode of collection - particularly the high usage of TDE and telephone in sub-annual surveys.

O Edit, imputation, outliers. Although many surveys will use similar approaches eg. software, there will be differences in edit rates and parameter settings.

O Rotation. Can differ between surveys and the rate of rotation will have an effect on the (standard errors of) estimates of change.

QUESTION 1: : Is the list of criteria for deciding when to benchmark correct?

2.2 Pre-requisites for benchmarking

Whatever the case for benchmarking a particular series, there are a number of simple pre-requisites. These are as follows:

C Annual variable concept is satisfactory. The annual survey must measure reasonably closely the desired concept.

C Annual reference period considered OK. Perhaps desirable, rather than strictly a prerequisite, it should be possible to align the annual reference period to that desired. As an example, ABI returns do not have to be calendar years, which is usually the required concept. However, it may be possible to adjust the ABI using the sub-annual profile - to "calendarise" it.

O Willingness to make revisions. By its nature, the process of benchmarking will cause revisions to sub-annual data. In some circumstances this can lead to several sets of revisions. Users will need to be convinced.

O Resources are available. The business area must have the resources required to calculate, check and publish the new estimates for the sub-annual survey.

QUESTION 2 Is the list of pre-requisites for benchmarking correct?

2.3 Some examples

The table below shows three examples.

Table 1. Illustration of the key criteria for Retail Sales Index, Capital Expenditure and Workforce Jobs

	RSI	Capex	Workforce jobs
Reasons to benchmark: (all vs ABI)			
Inferior variable concept	N	N	N
Inferior breakdowns (eg. variables or local units)	N	N	Y
Inferior estimation (eg. matched pairs)	N	N	Y
Inferior coverage of population	N	Y	Y
Inferior precision (incl response)	N	Y	Y
Inferior measurement at survey time	N	N	N
Inferior non-sampling errors	N	N	N
Consistency is paramount	N	N	Y
Pre-requirements to benchmark:			
ABI variable concept OK	Y	Y	Y
ABI reference period OK	N	N	Y
Willing to make revisions	Y?	Y	Y
Resources are available	Y?	Y	Y

This suggests a strong case for benchmarking Workforce Jobs, where matched pairs estimation is used at present, precision is inferior and consistency is paramount. WJ is in fact benchmarked. Meanwhile Capex has inferior coverage (small companies are excluded from the sub-annual sample) and inferior precision. Capex is also in fact benchmarked.

For RSI, there does not appear to be a major reason to benchmark to the ABI. RSI is not currently benchmarked although this is currently under examination.

3. A future ONS policy?

The conclusions of the thought process above are as follows:

- There should be no global policy to benchmark or not; each survey must be decided on its merits.
- There must be a good reason to benchmark. It is suggested that if none of the eight major criteria apply, there is no strong case to benchmark.
- Even if major reasons to benchmark exist, there are three or four prerequisites which need to be satisfied.
- Regardless of whether surveys are benchmarked or not, there is merit in the philosophy put forward by Stats Canada that wherever possible coherence should be built in at the design stage and that also there should be a process of annual reconciliation between sub-annual and annual surveys. This has resource implications.

QUESTION 3 Does GSSMAC concur that the ONS benchmarking policy should not be global ie it should be selective, judging each case on its merits?

References

Yung, Brisebois, Tardif, Kuromi and Rondeau. "Should Sub-Annual Surveys be Benchmarked to their Annual Counterparts? A Case Study of Manufacturing Surveys". (Stats Canada internal paper)

Office for
National Statistics

16th Meeting of the GSS Methodology Advisory Committee

Developing an Apportionment Method for Financial Variables Based on Returned and Synthetic Local Unit Turnover Data

Salah Merad, ONS

Executive summary

ONS business surveys collect data at enterprise level, and in order to produce regional estimates returns are apportioned between constituent local units. This is done via the application of a rule derived by fitting a model to single site and small multisite units. The rule is expressed in terms of local unit auxiliary information, including employment, economic activity and location. Concerns have been raised by users of regional statistics over a weakness in the apportionment methodology; in particular, there is a perception that the estimates for London tend to be positively biased in some economic sectors.

As part of the effort by ONS to improve the quality of regional estimates, a decision was made to collect local unit, or site, turnover data in the new Business Register and Employment Survey (BRES). The aim is to develop a new apportionment method that utilises the new data and to apply it to ONS surveys, ABI/2 in particular.

In 2008 we undertook a large pilot involving a random sample of about 11,000 businesses; the response rate has been good, around 80%. However, because of insufficient resources, we have been unable to validate the data. It was clear that some returns are not usable; an example is where a business apportions its turnover equally between all sites, irrespective of size or activity. To help us decide which returns are usable, we consulted the comments supplied by the respondents and devised simple rules. We found that about two thirds of the returns from multisite businesses are usable.

The first step in developing a new apportionment method involves predicting an August turnover value for every local unit in the register. This is done by fitting models to usable returned data and then applying the fitted models to all local units in the register. After analysing the collected data, we started fitting provisional models. It appears that Retail and Real Estate are different from other industries; they will need special attention.

In this paper, we have considered ways in which we can utilise the new site turnover data (returned or predicted) in the Annual Business Inquiry Part 2 (ABI/2). We present separate proposals for the apportionment of turnover and other financial variables. The proposed apportionment method for turnover is based heavily on predicted, or synthetic, site turnover and makes only a weak assumption. On the other hand, the proposal for other financial variables relies on the rather strong assumption that, conditional on register covariates and site turnover, local sites that compose multisite businesses are similar to single site businesses.

Aim of paper

- To present provisional models of local unit turnover and some proposals for the use of returned and synthetic site turnover data in the apportionment of ABI/2 financial variables.

Requested actions from the committee

- Feedback and guidance

Main issues for discussion

Question 1: Do you see any issues with the data we are using to build the models?

Question 2: Do you have any comments about the provisional models we have fitted, especially for Retail and Real Estate?

Question 3: Is the apportionment method we propose for ABI/2 turnover satisfactory?

Question 4: Would it be better to use synthetic BRES turnover for all units to produce annualized local unit turnover?

Question 5: Is the proposed apportionment method for ABI/2 financial variables other than turnover appropriate?

Question 6: Can we utilise information about the difference between single site units and local units to adjust the predicted values of local units for other financial variables?

Question 7: Could you suggest other apportionment methods?

Office for National Statistics

Developing an Apportionment Method for Financial Variables Based on Returned and Synthetic Local Unit Turnover Data

1. Introduction

ABI/2 collects data on a number of financial variables, including turnover and expenditure, from enterprises; many of these are composed of multiple sites that are geographically spread and/or active in different economic sectors. To produce regional estimates, the returned data from multisite enterprises are apportioned between constituent local units on the basis of a rule derived by fitting models to returns from single site or small multisite enterprises. The model covariates are based on local unit information, including economic activity (SIC code), region, and employment. It is thought that the resulting regional estimates are biased towards London, especially in some economic sectors such as Retail and Real Estate.

BRES, which will start in 2009, will collect local unit turnover for the month of August. In 2008 a large pilot was undertaken, where about 11,000 enterprises were sent a BRES questionnaire. We have now received responses from about 80% of businesses and the majority of the responses have been processed. We have started analysing the returned data and building a model to predict an August Turnover value for every local unit in the register.

In Section 2 we present an overview of the data received so far, and in Section 3 we present some provisional models we have fitted to the data. In Section 4 we give a quick overview of the current apportionment method used in ABI/2 and propose methods that make use of local unit turnover data. In Section 5 we summarise our findings and describe the next steps in this work.

2. Analysis of returned local unit turnover data from BRES

The BRES questionnaire requests information about the enterprise (Part A of the questionnaire) and information about each constituent local unit (Part B of the questionnaire). The questionnaire included a Part B questionnaire for every known constituent local unit and 3 extra questionnaires for any new local units; businesses can request extra Part B questionnaires.

A turnover question is included in Part B of the questionnaire, except in questionnaires sent to businesses in the public sector. Initially, we wanted to collect local site turnover for a whole year. However, because of concerns over the speed of availability of the data and the potential impact on burden and data quality, it was decided to request local site turnover for August, which is the month before the reference period of BRES. Here is the turnover question as it appeared it the 2008 pilot:

For the month of August what was the value of this site's turnover to the nearest pounds thousands (excluding VAT)?
Please round your figure to the nearest £thousand.

If turnover is nil, please enter a zero in the box
If exact figures are not available, please provide informed estimates
£_ _ _, _ _ _, 0 0 0

As of March 2009, we had responses from 7,892 businesses containing a total of 64,300 local units. Of the businesses that responded, 4,164 are single site enterprises and 3,728 are multisite enterprises. The local unit turnover returns can be classified into three groups:

- Positive value: 26,*490 local units*
- Zero value: 26,109 cases
- Missing: 11,701 cases

The number of local units with a 0 return seemed too high. We decided to examine the data in more detail.

2.1 Data quality

Because of insufficient resources, turnover has not been validated in the pilot. However, it was quite clear that many of the responses would not be usable for modelling. For example, a number of multisite businesses returned a positive value for one site and either a zero or a blank for all other sites. Some of these businesses included a comment in the available space saying that they couldn't split the business turnover between all their sites and so they lumped it all into one site. For units that included a comment it was quite straightforward to identify which returns to exclude. However, for units with no comment it was important that we applied some rules to exclude responses from multiple site businesses that were very likely to be in error.

Using information from the units with a comment, we identified as usable returns those that come from businesses with at least two sites with positive returns and where the proportion of sites with 0 or blank returns is not too high (less than 30%).

Multisite businesses with usable returns can be classified into 3 categories:

- Complete returns/comments: either all site returns are positive or the information in the comment indicates that the zeros are genuine (head office, warehouse). Some returned blanks, but because the comments are similar to where the return is zero, we think it's reasonable to assume that they should be zero.
- *Uncertain - excluding extreme cases*: no comment is available, but they have at least two sites with positive returns and not too many zeros or blanks.
- Special Arrangements (SA) - *excluding extreme cases*: it is composed of very large businesses that respond electronically (special arrangement cases), but with no comment box available. Businesses with too many zeros/blanks were excluded.

Table 1 provides information about the usable cases; it is complied from returns as of February 2009. We can see that about 2/3 of the multisite businesses provided usable data. The situation is quite good overall; however, for very large businesses it's quite poor. Only 12 out of the total 29 special arrangement businesses are deemed usable.

Type of response	No of businesses	Total no of sites	Total number of sites with zero turnover	Total number of sites with missing turnover	Proportion of zeros	Proportion of missing
Complete returns/comment	2,068	10,919	414	109	3.8%	1.0%
Uncertain -excluding extreme cases	189	4,119	290	149	7.0%	3.6%
SA - Excluding extreme cases	12	2,452	92	30	3.8%	1.2%
Total usable multisite Rus	**2,269**	**17,490**	**796**	**288**	**4.6%**	**1.6%**
All multisite Rus	**3,379**	**40,983**	**14,768**	**4,303**	**36.0%**	**10.5%**
Single site Rus	**3,763**	**3,763**	**172**	**334**	**4.6%**	**8.9%**

Table 1. Summary of returned site turnover data in BRES 2008 pilot

Zero returns

A zero return from a local unit can be valid; the returned questionnaires with comments indicate that this applies to a head office, or administration site, or a warehouse. Only 4.6% of the local sites constituting multisite businesses returned a zero value; another 1.6% returned a blank. As mentioned above, we think it's safe to assume that these are in fact zero values. Dead local sites are not included here.

The proportion of zeros in the single site businesses is also 4.6%, but the proportion of blanks is much higher (8.9%).

It is unclear whether the set of usable multisite businesses is unbiased in relation to the preponderance of zero turnover sites.

Region	Total number of sites	Number of zero turnover sites	% zero turnover sites
North East	590	13	2.2
North West	1,501	51	3.4
Yorkshire the Humber	1,269	37	2.9
East Midlands	1,026	36	3.5
West Midlands	1,708	53	3.1
East of England	1,289	49	3.8
London	1,841	87	4.7
South East	2,188	79	3.6
South West	1,340	61	4.6
Wales	731	21	2.9
Scotland	1,681	74	4.4
MISSING	2,325	235	10.1
Total	17,482	796	4.6

Table 2. Regional distribution of zero site turnover returns in BRES 2008 pilot

Table 2 gives information about the geographical distribution of zero turnover sites. The category 'MISSING' is composed of new local units for which the region code was not available at the time of this analysis. We can see that there is some variation, with parts of the North of England and Wales having a smaller proportion of zero turnover returns than the South of England and Scotland.

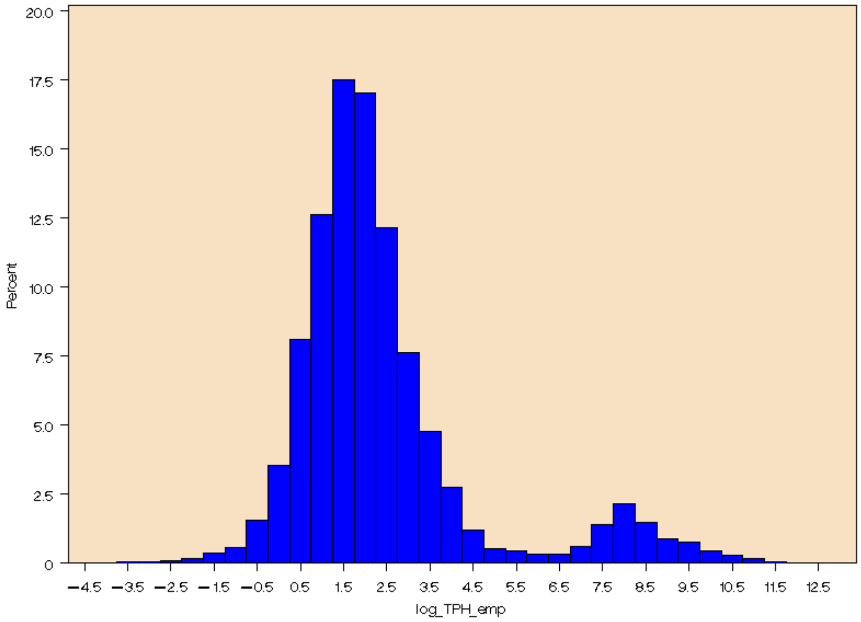
QUESTION 1: Do you see any issues with the data we are using to build the models?

3. Modelling August local unit turnover

The first aim is to obtain a turnover value for every local unit in the business register. The sample design of BRES means that large and complex businesses are selected every year, whereas medium businesses will be selected at least once in every four consecutive years. However, because of non-response and difficulties that businesses find in providing returns, we would still need to predict turnover values for the sites of some of the large and medium businesses.

We have considered modelling the zero values and positive values separately. In this paper, we only present models for the positive values. Given the skewness of turnover, we have considered the log transformation. Graph 1 shows the distribution of the logarithm of local unit turnover per head; it is clearly bimodal, indicating that there are different sub-populations.

An examination of the units on the right hand side of the histogram indicated that it is composed mostly of units from Retail (Division 47 under SIC2007) and Real Estate (Division 68 under SIC2007). There are also many units in which the returned local site turnover exceeds the annual turnover of the whole business; the latter is held on the business register. Most of these cases are very likely to be errors, and hence we have excluded them from the dataset used in modelling.



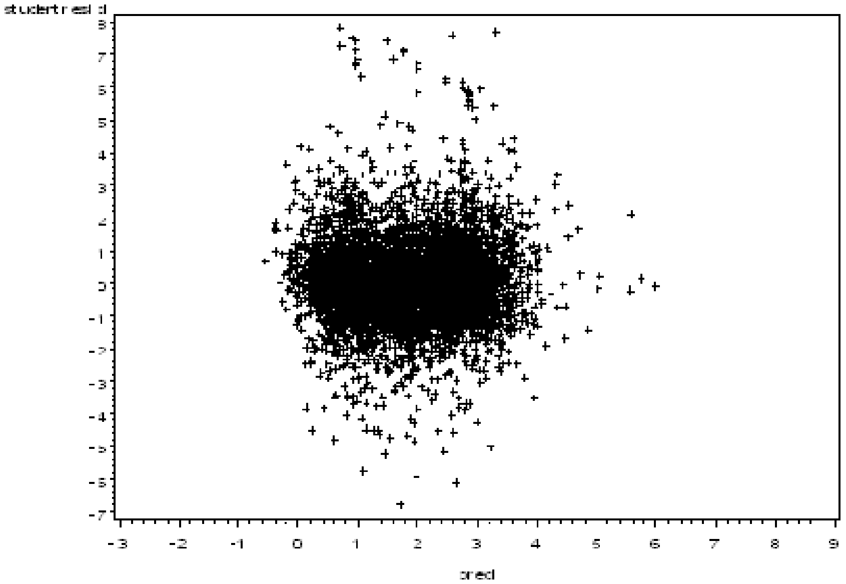
Graph 1. Distribution of log(site turnover per head) in BRES 2008 pilot

We have fitted an ANOVA model of log(site turnover per head) with the following as covariates: SIC, region, number of sites (categorised), employment band, enterprise register turnover

The studentised residuals plot of the model based on all positive usable data, excluding errors and data from Retail and Real Estate, is given in Graph 2. We can see that the pattern is quite random; the R^2 is 0.47.

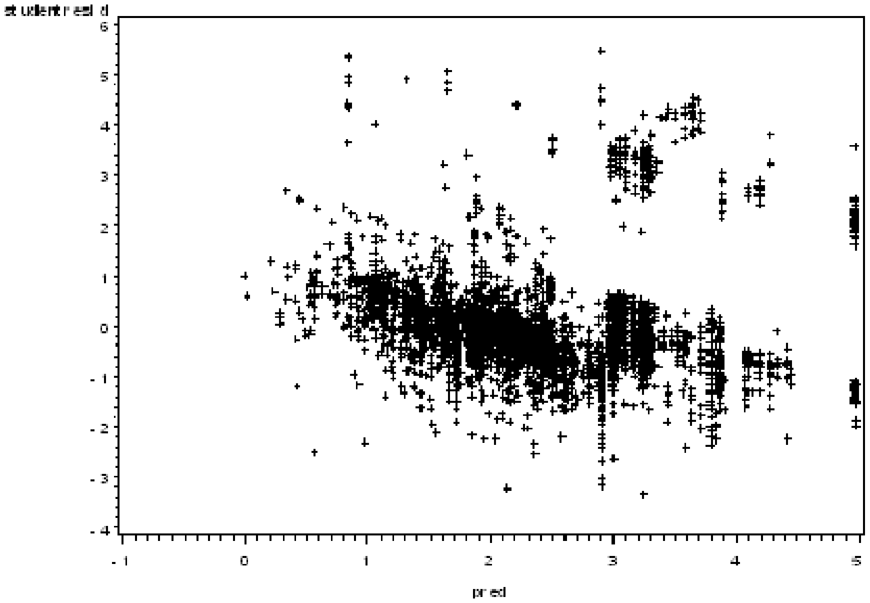
To check whether single site units are different from local units that are part of multisite units, we fitted a model that included a multiple site indicator and the register turnover per site as covariates, instead of the number of sites and enterprise register turnover. The coefficient for the multiple site indicator variable is found to be significantly different from 0. This indicates that turnover generating sites from multisite enterprises are potentially different from single site units for turnover. This could be an important aspect when we come to address apportionment of financial variables in ABI/2; this will be discussed in more detail in the next section.

We have also fitted a random effects model, to take into account any potential correlations between local units that belong to the same business. However, only a small number of enterprises have a random effect that is significantly different from 0. We need to look into this further to see if it a random effects model will add value.

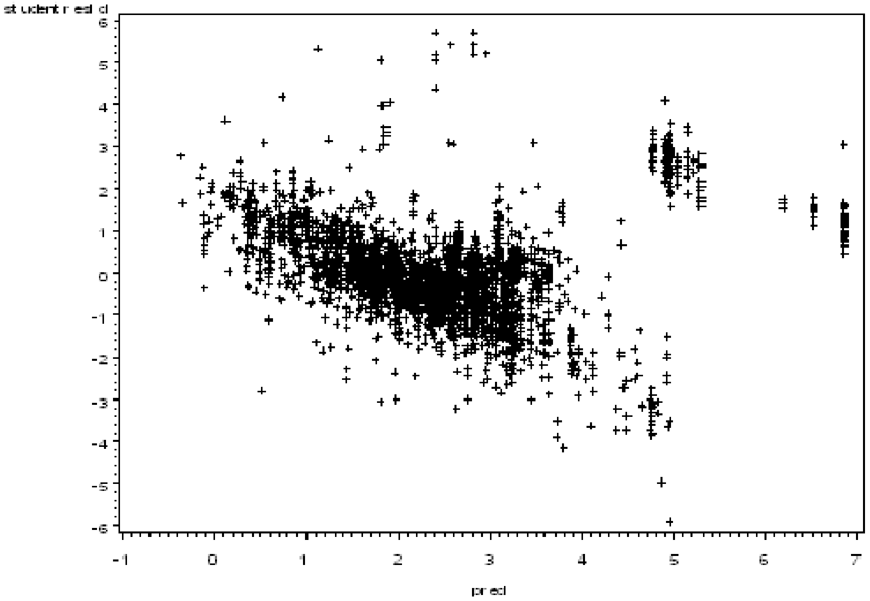


Graph 2. Studentised residuals plot of regression model – all industries except Retail and Real Estate

We have also fitted a similar model, without random effects, to units in Retail and Real Estate. Graph 3 shows the studentised residuals plot. There is a clear pattern in the plot, indicating that the model is unsatisfactory. Separating the two sectors did not make a difference, as the studentised residuals plot for Retail shows (Graph 4).



Graph 3. Studentised residuals plot of regression model for Retail and Real Estate



Graph 4. Studentised residuals plot of regression model for Retail

Summary

- Using log(site turnover per head) seems promising, with the exception of industries 47 and 68
- We need different models to deal with industries 47 and 68
- We need to model zeros – We are unsure about whether cases we identified as usable will allow us to obtain unbiased estimates
- We need to check the data further for possible errors.

QUESTION 2: Do you have any comments about the provisional models we have fitted, especially for Retail and Real Estate?

4. Apportionment methods based on local unit turnover: Some proposals

We start by giving a quick overview of the current apportionment method used in ABI/2.

4.1 Current apportionment method in ABI/2

ABI/2 data are used to produce regional estimates for a number of variables, including turnover, expenditure, stocks and GVA. The latter is derived from 11 ABI/2 variables, including the former variables. As returns are for whole enterprises, or reporting units, many of which are composed of sites that are spread geographically, it is necessary to apportion the returns between constituent sites. The apportionment is done via a model derived from single site and small multisite units. It is assumed that local units behave approximately like single site and small multisite enterprises. The model covariates used are employment band, SIC, and region. Because some of the returns are 0, the modelling is done in two stages:

Part 1: Fit a logistic regression to predict the probability that a unit returns a positive value for a given variable.
Part 2: Fit a regression model of log(survey variable/register employment) using positive returns only.

A separate model is fitted to each variable of interest.

Let y_i be the return by multisite enterprise i for survey variable y, and let $l = 1, 2, ..., L_i$, denote the sites that constitute enterprise i. The predicted value for local unit, or site, l, $\hat{y}_{il}$, is given by

$$\hat{y}_{il} = y_i \hat{p}_{il} \hat{y}_{il|y_{il}>0} \Big/ \sum_{l=1`}^{L_i} \hat{p}_{il} \hat{y}_{il|y_{il}>0} \text{ ,}$$

where $\hat{p}_{il}$ denotes the predicted probability that the value of variable y for local unit l in enterprise i is positive, and $\hat{y}_{il|y_{il}>0}$ denotes the predicted value of variable y for the local unit given that it is known to be positive.

Concerns have been raised by users of regional statistics over a perceived London bias, in particular in Retail and Real Estate. In an attempt to address this problem within the confines of the available data, ABI/2 and register data, we considered small alterations of the current models. These included modifying the definition of the set of units used to derive the model, adding interaction terms to the set of covariates and fitting separate models for different industries and regions. The regional estimates that resulted from each of the proposed models were very similar to the estimates obtained under the current models. For details of this work, see Merad (2008).

We concluded that the current apportionment method based on ABI/2 and register data was robust to small modifications of the models. We would need additional data at site level to address the concerns; BRES local unit turnover will provide the additional information that could improve the quality of regional estimates.

4.2 Proposals for the apportionment of ABI/2 variables using local unit turnover

In 2008 we collected local site turnover data and found that about two thirds of returned data would be usable towards producing a predicted site turnover for every local unit in the business register. Here, we present our initial thoughts about using the returned and predicted turnover local unit data to apportion enterprise returns between their constituent local units. We deal with the ABI/2 variable turnover and the other financial variables separately.

4.2.1 Apportioning ABI/2 turnover

ABI/2 collects enterprise turnover for a whole year, whereas BRES collects turnover for the month of August. Hence, before we can apportion the ABI/2 returns on the basis of BRES data, we need to put the BRES local unit values on an annual basis; we call this process annualisation.

Annualisation of BRES turnover:

To obtain an annualisation rule, one possibility is to use data from short term turnover surveys to model annual turnover, which will be derived from monthly returns, with the August turnover and other register data as covariates. Another possibility is to fit a model of ABI/2 turnover using single site enterprises, with August BRES turnover as one of the model covariates. As very few of the single site enterprises selected in ABI/2 would be selected in BRES, we would have to use the predicted August BRES turnover values. The two data sources, short term surveys and ABI/2, have different strengths and weaknesses with respect to the number of usable units and the coverage of industries. We will examine which is the most appropriate source in the near future.

Let $H\left(\mathbf{x}, z\right)$ denote the annualisation rule, where z denotes the BRES August turnover, returned or predicted, and $\mathbf{x}$ other business register variables such as industry classification and region. The derived annualisation rule H would then be applied to the local units that compose the multisite enterprises in ABI/2. The predicted annual turnover for local unit l in enterprise i, $\tilde{y}_{il}$ is given by

$$\tilde{y}_{il} = H\left(\mathbf{x}_{il}, z_{il}\right).$$

The apportionment rule becomes

$$\hat{y}_{il}^{(BRES)} = y_i \, \tilde{y}_{il} \Big/ \sum_{l=1`}^{L_i} \tilde{y}_i$$

Many of the multisite units in ABI/2 will have responses in BRES, and so we could use their returned values in the annualisation rule. For the other units we would have to use predicted values. These will all be positive, which means that sites that generate turnover will be slightly underestimated, whereas sites that do not generate turnover, such as head offices and warehouses, will be overestimated.

Note: for local units in which the response is a zero, the annualisation rule will not be suitable. The annualized value could be set to 0. However, there are cases where it is unknown whether the local units are head offices or warehouses. Some of these could be turnover generating units where for some reason the August turnover was 0. Using predicted, or synthetic, BRES turnover will circumvent this problem, but it is unclear whether this may introduce bias.

QUESTION 3: Is the apportionment method we propose for ABI/2 turnover satisfactory?

QUESTION 4: Would it be better to use synthetic BRES turnover for all units to produce annualized local unit turnover?

4.2.2 Apportionment of other financial variables

Apart from single site enterprises, we do not have site survey data for financial variables other than turnover, which we obtain using BRES. Hence, to apportion enterprise returns for financial variables other than turnover, we have to rely heavily on the model derived using single site data. This is similar to what happens in the current method but we could now use annual turnover as one of the model covariates.

Let $K\left(\mathbf{x},TO\right)$ be the rule derived after the two-stage modelling process of a given ABI/2 variable, other than turnover, where TO denotes ABI/2 annual turnover and $\mathbf{x}$ denotes variables from the business register.

When the derived rule K is applied to local sites that compose multisite enterprises in ABI/2, we will use the annualized site turnover values in place of ABI/2 returns.

This method should be satisfactory if the following assumption holds:
Conditional on the values of the covariates, including turnover, local sites from multisite enterprises are similar to single site enterprises with regard to other financial variables.

This may not be true and hence this method could introduce bias.

Local sites that do not generate turnover, such as warehouses and head offices, could well be dissimilar to single site units with regard to other financial variables. Hence, it would be inappropriate to apply the fitted model to such units. One way around this is to use annualized site turnover values based on predicted BRES turnover values. Local units with a predicted turnover value could be seen as quasi single site units where the predicted turnover is a financial proxy.

QUESTION 5: Is the proposed apportionment method for ABI/2 financial variables other than turnover appropriate?

We noted above that single site enterprises seem to be different from local units from multisite enterprises when modelling BRES turnover.

QUESTION 6: Can we utilise information about the difference between single site units and local units to adjust the predicted values of local units for other financial variables?

Remark: a simpler alternative to the proposed method would be to apportion all ABI/2 variables on the basis of annualised local unit turnover. However, when fitting ANOVA models of some financial variables, with ABI/2 turnover as one of the model covariates, we found industry classification and employment to be important predictors; the R^2 of the model with only ABI/2 turnover as the covariate is much lower. Hence, an apportionment that ignores the SIC codes and/or the employment of the sites could yield poor quality site values.

QUESTION 7: Could you suggest other apportionment methods?

5. Conclusion and next steps

In this paper, we gave an overview of the local site turnover data collected in the 2008 BRES pilot, and presented provisional models fitted to the data. We need to do more work to improve the models, in particular for Retail and Real Estate, including applying appropriate model diagnostics and validation.

We have also presented methods for using BRES site turnover data to apportion enterprise returns between their constituent local units. We have differentiated between turnover and other ABI/2 financial variables. We have noted that for the latter the method relies on the assumption of similarity between single sites and local sites. For turnover, we make the weaker assumption that the annualisation rule derived using single site data holds for local units composing multisite units.

A small proportion of local units returned a zero value, or a blank which could be taken as 0. We think that it would be more convenient to work with predicted values, rather than with returned values, especially with regard to financial variables other than turnover. Moreover, this means that all units, respondents, non-respondents and unselected, are treated in the same way.

We will be applying the proposed apportionment methods and/or other methods we may develop to 2008 ABI/2 data, and will compare the resulting estimates to the estimates obtained using the current method. When data from the full BRES are available in 2010, the models and the apportionment methodology will be finalised.

References

Merad, S. (2008) *Review of ABI/2 regional estimation methodology*, ONS technical report.

Office for
National Statistics

16th Meeting of the GSS Methodology Advisory Committee

Developing expertise in record linkage within ONS Methodology Directorate

Dick Heasman, Briony Eckstein, Peter Youens, ONS

Executive summary

ONS Methodology Directorate (MD) has implemented the plan, presented to the 12th meeting of NSMAC, to have one person full time working on record linkage. With further provision in MD's budget for people to work on this activity, we can now start to refer to a MD record linkage team.

One particular project the team is involved in is the sharing of data on school pupils with the Department for Schools and Families (DCSF). This provides an illustration of how shared data and linking such data can aid both National Statistics and research on pupil attainment, and also demonstrates the procedures necessary for data sharing across Government to take place.

Aim of paper

The aim of this paper is to outline early initiatives in the area dedicated to developing capacity for record linkage within ONS, using one specific example of a project involving data sharing with another Government department as an illustration, and to present plans, on which the committee is invited to comment, for how the work will develop.

Requested actions from the committee

The committee is asked to take note of the opportunities for, and constraints on, sharing data across the GSS, and to comment on MD's plans for developing the capacity for record linkage within ONS. The committee is asked to advise us on other centres of expertise in record linkage in the UK with whom we might establish useful links.

Main issues for discussion

The GSS MAC is asked to comment on MD's plans for the development of the work on record linkage.

Developing expertise in record linkage within ONS Methodology Directorate

1. Introduction

The use of data from administrative sources has the potential to deliver statistical and analytical benefits including a reduction in the data collection burden, more accurate statistics and improved policy analysis. With this in mind, a paper entitled *Combining Data: Developing a Centre in MD to meet the challenges* was presented to the 12th meeting of the NSMAC on 11 May 2007. The paper and the committee conclusions can be found at:
http://www.statistics.gov.uk/methods_quality/nsmac_twelfth_meeting.asp .
Section 3 of the paper outlined ONS plans to carry out a project for the Department for Transport (DfT) in 2007/08 as well as to develop work in record linkage more generally. The post of one person full time to carry out this work was filled between July 2007 and April 2008 and resulted in the completion of the project as well as the gathering of a considerable amount of technical expertise that has been handed on in the form of, for instance, computer programs, papers and presentations from conferences. The linking work for DfT involved close collaboration between DfT and ONS, and was presented jointly at the 2008 GSS Methodology Conference. The conference paper can be seen at: http://www.ons.gov.uk/about/newsroom/events/thirteenth-gss-methodology-conference--23-june-2008/programme/index.html , (session 6 at 14:00).

MD continues to dedicate resource to enhancing capacity for record linkage within ONS and the wider GSS, with the posts allocated to the Small Area Estimation Centre. We are now in a position to start referring to the MD record linkage team, making this an appropriate time to make an approach to GSS MAC for guidance on the team's work.

The work in MD will focus on developing technical expertise in record linkage. Although the MD record linkage team will be aware of the ethical and data protection issues involved, and the legislative framework, detailed work on these issues will be carried out in other branches of the ONS, notably the Administrative Data Development Team, the Legal Services Branch and the Statistical Disclosure Control Branch. Similarly, the analysis of linked datasets would generally be carried out by the areas with direct need for such enhanced data, such as the ONS Centre for Demography (ONSCD). MD would, however, see part of its role as advising analysis centres on the issues and difficulties involved in analysing linked datasets.

QUESTION 1: Does the committee have any comments on the issues and difficulties involved in analysing linked datasets?

This is one example of where the team would see its role as being broader than just developing expertise in the linkage process itself. Denk (2008) gives a useful categorisation of the Record Linkage - she calls it Statistical Entity Identification (SEI) - Framework as Preparation (parsing, standardisation, phonetic coding), Candidate Selection (e.g. blocking), Comparison, Scoring & Classification, Decision, and Evaluation (the estimation of quality measures). To these we would add at the beginning the pre-linkage processes of variable selection, formatting and quality checking and at the end, logical checking of the resulting linked datasets and advising on the use of them in further analysis.

QUESTION 2: Does the committee think it reasonable for MD to concentrate on developing expertise in this set of processes?

The rest of this paper is set out as follows: section 2 discusses the background to submitting the paper, particularly how it has changed since May 2007. Section 3 discusses the role of the MD record linkage team in providing learning and development opportunities, to ONS staff initially. Section 4 and 5 are about how expertise on record linkage can be shared and disseminated across ONS staff and GSS staff respectively. Section 6 is about the projects currently under way or in preparation where the team will take on an advisory role. This is divided into subsections, with subsection 6.1 going into much more detail than the rest of the paper describing projects to share data with DCSF for record linkage purposes. Finally section 7 deals with the future plans of the MD record linkage team.

2. Background

The context has developed somewhat since Cruddas (2007). Major events were the passing of the Statistics and Registration Service Act in July 2007, and it coming into force in April 2008. The Act provided for secondary legislation (regulations) to overcome legal barriers to data sharing between public authorities and the UK Statistics Authority for statistical and analytical purposes.

ONS has become involved in many more record linkage projects, for example one to link the Annual Population Survey (APS) database to Individual Learner Record data. MD has also become aware of the amount of record linkage taking place in other government departments, and although some of these projects (e.g. in HMRC) are aimed at gaining intelligence rather than having a statistical purpose, there is much experience which can be shared across the GSS. The need to carry out the matching recommended under the interdepartmental Task Force on Migration Statistics (December 2006) has now become urgent and ONSCD need advice and practical help with linkage methods.

GSS has had a Data Sharing subgroup (of the GSS Statistical Policy and Standards Committee) since March 2007, while ONS has had a Data Sharing Steering Group since September 2008. Annex A shows the combined March 2009 progress report from these two groups, to give a flavour of the data sharing activity taking place.

3. Learning and development

The MD record linkage team sees the provision of learning and development activities as one of its more important functions. It has started by building a library of useful references in the form of books, presentations and papers. It has also organised a training course for ONS staff entitled ‘Record Linkage – Theory to Practice’ taught by Dr Natalie Schlomo from Southampton University, and attended by sixteen participants. There is a demand for more courses of this type, particularly among ONSCD and Census Directorate staff.

The team is also promoting the courses run by the ADMIN[2] node of the ESRC National Centre for Research Methods, both among ONS and wider GSS staff.

QUESTION 3: Can the committee advise on other centres of expertise in record linkage in the UK with whom we might establish useful links?

Both during the 2007-2008 period and in the current period MD record linkage team members have attended suitable conferences and workshops, and documentation from these events has been or will be made available. The team is on the mailing list of the "Integration of surveys and administrative data" group of the European Statistical System (ESS).

QUESTION 4: Can the committee suggest suitable Conferences to attend, or sources of training provision for record linkage other than those mentioned in this section?

For MD plans to deliver learning and development, see section 7. Future plans.

4. Sharing expertise within ONS

The MD record linkage team is compiling a directory of ONS staff who practice or have experience of record linkage. There appears to be a significant amount of practical experience of record linkage already in the ONS, though not necessarily all residing within a single individual or team. Therefore the directory will be used as a basis for setting up forums and workshops (see also future plans) within ONS, where knowledge and experience can be shared. At the same time, it should be remembered that some of this experience may have been gained several years ago, so part of MD’s role is to keep abreast of latest developments in the field and to use such events to disseminate them.

The team has set up a common folder on the ONS network for record linkage practitioners to share useful examples of code. Two examples are code to create missing value flags and code to implement the NYSIIS system for name coding. Most but not all of these codes are in the SAS software language.

5. Sharing expertise with the wider GSS

The MD record linkage team has been building up contacts with practitioners of record linkage in other government departments. In the past, staff at HM Revenue and Customs (HMRC) have organised a cross-departmental data matching forum, though this has not met for a year now. The aim is to add some impetus to get it meeting regularly again.

[2] ADMIN (Administrative data: methods inference & network) aims to exploit newly linked administrative and survey longitudinal data, to develop and disseminate methodology for making the best use of administrative data and to reassess how best to deal with some of the common problems associated with using survey based longitudinal data. For webpage see references.

The interaction between the MD team and other government departments can be seen as very much a two-way process. For instance, two staff from MD recently managed to attend a meeting set up internally for HMRC and were able to make useful contacts and gain from that department's considerable experience in the field, while in dealings with the Department for Children, Schools and Families (DCSF) it is mainly ONS staff who take on the advisory role.

6. Advisory role on projects involving shared data and record linkage

One of the major roles of the MD record linkage team is to advise on projects involving shared data and record linkage. This could involve actually doing the linking on behalf of other business areas in the ONS, particularly in projects taking place in the next year or so as this should lead to the team gaining valuable experience and expertise. Examples follow of the projects currently under way (although in their early stages).

6.1 Sharing data with DCSF

Good quality population and migration statistics are essential for providing the evidence base for managing the UK economy, planning and allocating resources. Improving the quality and range of these statistics has been a key priority for ONS and good progress has been made in the past two years. Key developments include improved regional and local estimates of migration, short-term migration estimates, and cross-government migration reporting. A programme of further improvements is being taken forward by ONS and other government departments. These initiatives include further improvements to statistical methods, development of more timely indicators, and exploitation of survey and administrative data sources.

The recommendations set out in the report of the Interdepartmental Task Force on Migration highlighted the importance of gaining access to data from a number of administrative sources including the School Census. The GSS Data Sharing Steering Subgroup agreed that data sharing to support improvements to migration statistics is a top priority.

Draft regulations for sharing data must be supported by evidence showing how each data item would be used. A joint feasibility study was therefore agreed between ONS and DCSF to research issues related to data sharing. The study has two strands addressing 1) ONS's need for data from the School Census, and 2) DCSF's need for survey data from ONS. The first phase of this work was designed to gain a better understanding of the scope and content of specific sources including the School Census and the National Pupil Database (NPD) by collaborative working between relevant staff from ONS and DCSF. To take this forward, two members of ONS staff worked with the full cooperation of staff from DCSF Data Services Group in Darlington for a week during March 2008.

This first phase led to a good level of familiarity with the strengths, weaknesses and level of coverage of the School Census and NPD. The overall conclusion from the analysis was that the quality of the School Census is high. In addition, it was concluded that further work should enable ONS to have, for example, additional information from the School

Census on the population aged 5-15 including numbers broken down by language and ethnicity, and a better basis for developing and implementing methodological improvements to the mid-year population estimates process. DCSF indicated that it would welcome support from ONS for the proposed inclusion of a migrant flag or indicator in the School Census from 2010.

6.1.1 Linking School Census data in pursuit of better population and migration estimates

Record level data from the School Census covers maintained schools in England and provides information on approximately 7 million children. This information should be of great benefit to ONSCD as it has the potential to contribute to the validation and improvement of existing population estimates. At the same time accumulation of these records over time will enable overall changes to the school age population to be monitored alongside variations in the number and characteristics of migrant children and their movements. Such data will contribute to a better understanding of key components of local population change, provide vital evidence about internal migration flows and support ongoing work to improve small area ethnic estimates.

The pupil level data provided by the School Census gives individual level information on all children within the English maintained school system. Variables of particular interest include:

- Unique Pupil Number
- Pupil's full name (and any previously held names)
- Pupil's date of birth
- Sex
- Whether English is the pupil's first language
- Pupil's ethnic group
- Pupil's full residential address
- Dates of entry into and out of the pupil's current school
- Whether the child is a boarding pupil
- Whether the child has a parent serving in the home armed forces
- School identifier

Some of these variables, such as Unique Pupil Number, Name, and Date of Birth are primarily of use in linking information over time within the School Census, and in linking to other sources. Variables such as ethnicity and language will inform on the characteristics of both stocks and flows of the school age population, with this information being available at quite low levels of geography due to residential address information (subject to it not being disclosive).

By linking data from the School Census to other sources (e.g. surveys, other administrative data or population censuses) it should be possible to obtain statistical information not readily available elsewhere about child migrants, children of migrants and, by association, their parents. For example, combining information on age, sex, previous and current address from the School Census with data from surveys and other sources should provide better local information on the numbers of families with young children.

Such information has the potential to contribute to methodological developments and improvements designed to improve the accuracy of population estimates and provide a more reliable basis for the determining the assumptions which underpin longer term population projections. These considerations formed part of the case put to Parliament in support of the data sharing regulations.

Initially the project will attempt to link record level School Census data from children aged 5 and 6 years in the January 2008 collection to a sample of births data[3] from 2002. As well as enabling the development of expertise in linkage within ONS, this work will allow ONSCD to develop a better understanding of local population change and establish whether or not there are differences for those whose birth records can, and cannot, be successfully linked with information from the School Census. It is also anticipated that this work will aid understanding of the discrepancy between the number of births registered in England in the year prior to the 2001 Census and the number of under ones enumerated at the 2001 Census. Comparison of the demographic characteristics of those cases that can be linked with those unlinked will inform on those that are harder to track. This information should help in understanding the demographic characteristics of those under ones not identified in the 2001 Census. For example:

- are certain ethnic groups more likely to be linked?
- Using birth information on mother's country of birth, is there a relationship between this information and whether a child is tracked?
- Using information on location of maintained and independent schools, are those linked more likely to be in areas where a high proportion of the school age population attend maintained schools – the default then being that those not tracked are more likely to be in areas where there is a lower propensity to attend maintained schools.

Using the skills developed and knowledge gained by linking the subset of record level School Census data to birth records, the intention is then to link the full dataset to NHS General Practice registration data. The aim is to enhance knowledge of migration (in particular internal but also international) and transform this knowledge into improvements in migration statistics.

QUESTION 5: Can the committee comment on the ways the data shared between the ONS and DCSF can be used to improve migration statistics?

The work on linking the School Census data and the 2002 births data sample has now started, with both having been successfully loaded into the Virtual Microdata Laboratory (VML) and transferred into SAS. The first stage is to make initial quality checks of both sources, which comprise:

- Assessing the number of missing values
- Checking for plausible sex ratios
- Taking the five most popular names for each sex, and checking them against sex
- Plotting dates of birth to look for any abnormal peaks or troughs

[3] 36 dates of birth in 2002 have been selected, with the aim of linking using the variables Surname, First Name, Middle name(s), Sex and Date of birth

- Searching for duplicates on matching variables
- Checking that the postcode variables can be mapped to Local Authority (LA)
- Ensuring there is a look up table to match the data to LA

The next stage is to ensure common formats across both sources, to include dates of birth, name and address fields and to ensure common coding systems for the geographic variables.

It is then proposed to spend some time working with the 2008 School Census data only. Aggregating it by single year of age and other variables such as Local Authority and ethnicity will aid the validation of 2007 mid-year population estimates.

It is proposed to link the School Census data and the births data sample, matching on sex, date of birth and name, as follows:

1. Reduce the School Census data set to those with the same date of birth as in the births data sample.
2. Block the data by sex and date of birth.
3. Do an exact match by name. Resolve any duplicates. Record the number of links by those that have the same LA in the School Census as at birth and those where it differs.
4. On the unmatched residue, use a name coding system and match by name. Clerically check a sample of the links produced. Then proceed as in step 2.
5. On the unmatched residue, block by sex and region or LA.
6. Do an exact match by name.
7. Clerically examine the links produced and keep those where the discrepancy in date of birth can plausibly be put down to recording error.
8. Calculate the percentage of cases matched from the births sample and subdivide this into cases with the same or different LA and cases with the same or different date of birth.

It is recognised that the use of middle names might be a problem, and this scheme might have to be modified to account for middle names being left out of the School Census.

6.1.2 Using School Census data to improve the 2011 Census

The other part of the business case supporting the data sharing regulations was that access to the School Census data would be useful for two purposes in the 2011 Census of population:

i) developing, planning and implementing effective enumeration strategies to improve response rates
ii) assessing coverage and making adjustments to improve the 2011 Census.

The first purpose uses only aggregate data. The data will be aggregated to LSOA (Lower level Super Output Area) to give a timely indicator of variable values, such as certain ethnicities and English not being the first language, associated with hard-to-count areas.

For the second purpose, data linkage of School Census data is being considered to potentially inform Census Quality Assurance. There are two purposes for which data linkage might be undertaken:-

1. To assess the coverage and completeness of School Census data to evaluate its suitability and inform its use as a Census comparator at aggregate level.
2. Along with a range of other administrative sources, data linkage of School Census data is being considered as a means of filling information gaps identified through Census validation.

QUESTION 6: Can the committee comment on the ways the data shared between the ONS and DCSF can be used to improve and validate the results of the 2011 Census?

6.1.3 Obtaining the School Census data and an overview of the legislative process

The Statistics and Registration Service Act 2007

The data sharing powers of the Statistics and Registration Service Act 2007 (the 2007 Act) enable regulations to be made that remove legal bars to data sharing between the UK Statistics Authority and other public authorities for statistical purposes. Such regulations require the consent of the Ministers involved and approval by Parliament. Regulations made under the 2007 Act cannot override human rights or data protection legislation. Any data sharing proposal must be compatible with the European Convention on Human Rights.

Initial Legal Assessment

The 2007 Act only allows for regulations to be made where there is a legal bar or there is no legal authority for the data to be shared. An essential first step was therefore to establish whether a legal gateway existed that would allow the data owners (DCSF) to share the data with ONS for the statistical purpose. Lawyers representing ONS, DCSF and the Cabinet Office (CO) carried out legal assessments which concluded no such legal gateway existed.

Feasibility work

The work with DCSF in Spring 2008 described in section 6.1 provided the basis for understanding key concepts and data definitions and contributed to a better understanding of the scope, content and quality of the data. It helped to identify those data items needed to meet specific statistical requirements, and provided an understanding of data collection processes including the steps taken to clean and validate the data. This preliminary work helped to substantiate the requirements for record-level data. The collaborative approach pursued by ONS and DCSF continued throughout the process to obtain the regulations and was key to its success.

Preparation of business case

Section 47(9) of the 2007 Act requires that any proposal to share data must be supported by explicit evidence to show how the information will be used, and to explain why aggregate outputs cannot meet the statistical need. The knowledge and understanding gained from the feasibility work enabled the identification and selection of relevant data items. The ONS Administrative Data Development Team developed the business case in consultation with colleagues from the analytical areas which required the data. The business case provided a detailed explanation of how individual data items would be used.

Engagement with Cabinet Office; submissions to Ministers

The CO retained residual responsibility for ONS following its independence in 2007. The CO is responsible for drafting and laying regulations to be made under the 2007 Act. Based on the draft business case submissions were made by CO officials to the Minister (MCO) and by officials in DCSF to the Secretary of State (SoS) in June 2008. These informed them of the proposed data sharing and asked for permission to proceed. A final submission was made once drafting of the regulations had been completed seeking approval to lay them before Parliament.

Recent high-profile data losses have underlined the need for transparency and the importance of ensuring appropriate steps are taken to safeguard the transfer and handling of personal data. In response to Ministerial concerns on data security ONS Legal Services branch set out ONS's commitment to ensuring the security and confidentiality of data in its possession and the arrangements that would be put in place for the transfer, storage, handling and access to data from the School Census. These take full account of the recommendations of the Government's Data Handling Review and other related requirements for secure data handling.

The business case provided a comprehensive description of the statistical requirements and a detailed explanation of how each data item would be used. It formed the basis for CO lawyers to draft the Regulations. ONS dealt with lawyers' detailed queries regarding the proposed use of and statistical justification for data items.

The draft regulations were laid before Parliament and published in early December 2008. ONS and DCSF had worked co-operatively prior to this to produce coordinated and coherent press handling strategies and high-level lines to take to deal with potential controversies that might have been raised by the press or pressure groups. Following scrutiny by the Lords Merits of Statutory Instruments Committee ONS provided additional information on consultation that had taken place with stakeholders.

Parliamentary debates

The draft regulations were debated in the House of Commons in late January 2009 and in the Lords a week later. Officials from ONS produced comprehensive briefing covering potential questions for the MCO prior to the debates. Officials from ONS and DCSF were present to support the MCO during the debate in the Commons. Following approval by both Houses and signing by both Ministers the regulations came into force on 11 February.

Departmental Processes

Disclosure of the data to ONS additionally required clearance through DCSF's internal processes. ONS co-operated with DCSF to ensure that the latter's specific requirements relating to data transfer, security and confidentiality were met. Following satisfactory conclusion of the arrangements DCSF disclosed the data to ONS on 1 April.

Data Access, Storage and Handling

A secure pilot research environment for administrative microdata was established within the Virtual Microdata Laboratory (VML). Security features of the VML prevent data being transferred into or out of the facility other than by authorised administrators who control access and who are responsible for monitoring the use of the data. Access to School Census data is restricted to staff working directly on improvements to population and migration statistics. As part of the arrangements agreed with DCSF, ONS carried out Criminal Records Bureau checks on staff before granting access. These staff were given relevant training and made aware of their responsibilities for maintaining data security and confidentiality and the penalties under the 2007 Act for disclosing or sharing information unlawfully. The Regulations do not permit ONS to onwardly disclose the data. ONS has given an undertaking to the Secretary of State in DCSF that it will obtain the consent of DCSF before disclosing data to contractors.

6.1.4 Pupil attainment and the Labour Force Survey (LFS)

The Labour Force Survey is a rich source of data which include variables on (among other topics) individual demographics, household and family characteristics, economic activity, employment details, unemployment duration, education, training and income. DCSF have asked for access to this data source, for the years 2004 to the most recent available, to carry out research on the associations between selected variables on these topics and pupil attainment as recorded in the National Pupil Database. The NPD is thought to contain data even more sensitive than in the School Census, so DCSF asked ONS to prepare and share with them LFS data in order to carry out this exercise. It was agreed that the most comprehensive source of LFS data that could be used was the Annual Population Survey (APS), which is in effect the largest possible annualised data set of unique individuals in the LFS.

There are no variables on the publicly available APS data sets held at the National Data Archive that could provide a sufficiently accurate link to the NPD. The primary sampling unit of the LFS is address, and, subject to approval being granted by the Microdata Release Panel, ONS agreed to supply the address data attached to the microdata, which includes date of birth. The NPD covers schools in England only, but ONS proposes to supply LFS cases for Wales and Southern Scotland as well, to cover any cross-border catchment areas.

At the time of writing this paper, the list of variables to be supplied and the business case for the transmission of the data to DCSF are being prepared. Annex B shows the latest proposed list of LFS variables to be included, but this may still change.

When completed, the business case will be submitted to the Microdata Release Panel. In readiness for this, the record linkage team has been investigating the reconstitution of the APS records with the address data. It has found that this will be possible using address data supplied by the Sampling Implementation Unit of the ONS with addresses that entered the sample back to 2003. This means that ONS will be able to supply addresses for all the required cases in the 2007 APS, and a progressively smaller proportion (but still a clear majority) as we go back to the year 2004. Reconstituted 2008 APS data can also be supplied when the 2008 APS microdata becomes available. Proof of concept for the reconstitution process has been demonstrated by matching address variables to the spine of the 2007 APS file and testing their plausibility.

No detailed linking strategy has yet been devised for linking the APS data to the NPD data if approval from the Microdata Release Panel is forthcoming. It is envisaged that the data sets will be held in DCSF's own secure data facility, that the linking work will be carried out by DCSF staff (not contractors) and that members of the MD record linkage team will act as advisors when required. The matching would be done primarily on sex, date of birth and address.

QUESTION 7: Do the committee wish to comment on any aspect of the DCSF project to research pupil attainment as outlined in this subsection?

6.2 Sharing Migrant Worker Scan data

The Migrant Worker Scan (MWS) is a data set created by HMRC from all applicants for a National Insurance Number (NINo), whether for work or benefit purposes, who are migrants to the UK. Some migrants are not captured, for example children, migrants who choose not to work, and, since the definition of a migrant includes persons who previously had a UK NINo but who were subsequently resident abroad for a period of more than a year, migrants previously allocated a NINo while they were UK residents. However, it will include the vast majority of foreign-born migrant workers. Age, sex, postcode of residence, country of origin and year of arrival are among the more useful variables recorded from the point of view of improving migration estimates.

HMRC takes a quarterly snapshot of this database and passes it to the Department for Work and Pensions (DWP), who subject the data to a thorough cleaning process before producing their own National Statistics from them. Deaths of people still resident in the UK are recorded, but people who emigrate are unlikely to be removed from the list. The database therefore tends to be an ever-expanding list. These data cannot be used in isolation to infer migration flows or stocks of migrants resident in the UK, but have potential when used in combination with other sources.

There is a legal gateway for the MWS to be shared between DWP and ONS. At present ONSCD are working with three snapshots: to September 2007; to December 2007; and to June 2008. In the short to medium term their work is in two stages. The first stage is linking data in the three snapshots to assess how many migrants change their address over time and where they move to. The MWS will include encrypted NINo to facilitate this stage.

The second stage is due to start shortly and is where it is hoped the MD record linkage team will provide a significant input. This will be linking MWS snapshots to GP patient registers to see where migrants settle. If it is possible to link records in the MWS to patient registers, it will be possible to track migrant moves. However, the unique identifiers on each dataset are different - encrypted NINo on the MWS and NHS number on the patient registers. Neither the DWP extract of the MWS or the patient registers currently held in ONS holds name and therefore matching would need to be carried out using date of birth, sex and postcode. At this stage, it is unclear whether there are sufficient variables to successfully match.

7. Future plans for the MD record linkage team

Core function The team aspires to become first port of call within ONS for advice on record linkage. This is dependent on the team's learning, and its growing experience of data linkage projects. It also depends on the team's success in carrying out its role in the projects discussed in section 6. The team will publicise its role, initially to the rest of ONS Methodology Directorate and then more widely, through for instance presentations at seminars and a presentation at the 2009 GSS Methodology Conference.

Record linkage projects The MD record linkage team plans to continue its advice and support to the project involving Migrant Worker Scan data into two further stages of the research that ONSCD propose to undertake. These are described briefly in Annex B. Support is also planned for linkage work connected to the Beyond 2011 Project and to the Annual Survey of Hours and Earnings, and to a review of the working of the Virtual Microdata Laboratory.

Learning provision In the current year the team intends to organise forums for ONS staff to share experience and good practice, and to contribute to the delivery of the Statistical Analyst Module on Administrative Data and Data Linkage, which comprises a day in the classroom followed by an assignment. In the longer term, the team plans by 2010/2011 to provide a full one-day Methodology Workshop on record linkage and by 2011/12 to be able to help to meet the need for training and learning in record linkage in the wider GSS.

Software At present it is assumed that most matching projects in ONS will be carried out using SAS. The team propose to do an evaluation of this software and others, including commercial data matching softwares, for their capabilities in the field of record linkage. In addition, the team would wish to investigate the capabilities of the implemented components of the SEI framework suite of R programs (Denk 2008).

Statistical matching At present the attention and activities of the team are confined to one-to-one matching, both deterministic and probabilistic, and to striving for expertise in these fields. Statistical matching, or data fusion, aims to integrate data sets that contain information on a set of common variables where the entities in the data sets are different. The team plans in the medium term to conduct research into the potential value of statistical matching for Official Statistics in the UK, for instance in providing a type of synthetic data set that can be used to improve estimation methods. The aim is to carry out this research in 2010/11.

Standards and Guidance A longer term plan for the team is the production of a set of standards and guidance for record linkage. Realistically it is unlikely that this work would start before 2011/12.

GSS Methodology Series paper Gill (2001) is the GSS Methodology Series monograph on Methods for Automatic Record Matching and Linkage and their Use in National Statistics. Another longer term aim for the team is to produce an updated monograph for the series.

QUESTION 8: Can the committee comment on the future plans of the MD record linkage team as outlined in this section?

References

The webpage for ADMIN is at
http://www.ncrm.ac.uk/about/organisation/Nodes/ADMIN/

Cruddas, M., (2007).
Combining Data: Developing a Centre in MD to meet the challenges
Paper for the 12th meeting of NSMAC, 11 May 2007. Paper and committee conclusions at http://www.statistics.gov.uk/methods_quality/nsmac_twelfth_meeting.asp

Denk, M., (2008).
A Framework for Statistical Entity Identification to Enhance Data Quality
Paper from the proceedings of the CENEX project workshop 29-30 May 2008. http://cenex-isad.istat.it/ navigate via Announcements → Workshop → Workshop papers → session4 1.pdf. If you have problems, email dick.heasman@ons.gsi.gov.uk for a document attachment.

D'Orazio, M., Di Zio, M. and Scanu, M., (2006)
Statistical Matching: Theory and Practice
Wiley: ISBN: 978-0-470-02353-2

Gill, L., (2001).
Methods for Automatic Record Matching and Linkage and their Use in National Statistics
Monograph for GSS Methodology Series NSMS25.
http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=9224

Herzog, T. N., Scheuren, F. J. and Winkler, W. E. (2007) .
Data Quality and Record Linkage Techniques
New York: Springer. ISBN 978-0-387-69502-0.

Annex B: Suggested 2007 LFS variables to be used in research on pupil attainment

Variable type	Variable name	Variable description
System variables	CASENO	unique case identifier
	ADDNO	unique address identifier, derived by MD team
	HHLD	household number at address
	FAMUNIT	family number within household
Individual demographics	ADFIELD1	Address, part 1
	ADFIELD2	Address, part 2
	ADFIELD3	Address, part 3
	ADFIELD4	Address, part 4
	DISTRICT	Address, part 5
	POSTTOWN	Address, part 6
	POSTCODE	Address, part 7
	SEX	Sex of respondent
	DTEOFBTH	Date of birth
	ETHCEN15	Ethnicity revised
	CAIND	Child/Adult indicator
	GOVTOF	Government Office Region
	UALADGB	Unitary Authority/Local Authority District
Household characteristics	HOHID	Head of household (indicator)
	TEN1	Accommodation details
Economic activity	INECAC05	Basic economic activity (ILO definition) (reported)
	ILODEFR	Basic economic activity (ILO definition) (reported)
	SCHM04	Government employment and training programme
Main job	INDD92M	Industry class in main job
	SC2KMMJ	Major occupation group (main job)
	NSECM	NS-SEC category (main job)
	NSECMMJ	NS-SEC class (main job)
ILO unemployment	DURUN2	Duration of unemployment
	BENFTS	Whether claiming any State Benefits/ Tax Credits
Education and Training	HIQUAL5	Highest qualification/ trade apprenticeship
	HIQUAL5D	Highest qualification (detailed grouping)
	QULHI4	Highest qualification currently studying towards
	CURED	Current education received
Income*	GRSSWK	Gross weekly pay in main job

*The MD team has advised DCSF that ONS does not consider the LFS to be the best survey source for income data. However, the inclusion of another survey source would be costly and delay the project.

Table 1: Suggested 2007 LFS variables to be used in research on pupil attainment

Annex C: Stages 3 and 4 of the matching project using Migrant Worker Scan data

Stage 3 The extract of MWS data to be received from DWP does not contain name - this would only be available via data from HMRC which would involve delays to the timetable and would be likely to have a cost associated with getting the data. Linkage without name may prove problematic. One solution could be to link the MWS to the Lifetime Labour Market Database (LLMDB) which is a 1% sample of the National Insurance Recording System, using the encrypted NINo. The LLMDB does contain name and so it would allow us to assign a name to 1% of the MWS data. ONSCD could then assess whether 1% of the MWS data could be linked more successfully to the patient register if name were available. If name was deemed necessary, future extracts of the MWS would have to be requested from HMRC.

Stage 4 ONS is not likely to have access to the Work and Pensions Longitudinal Study (WPLS) for some time yet (definitely not until the second half of this year) but when it does it would like to link MWS snapshots to WPLS data to try to track migrant moves. The MWS holds encrypted NINo and it should be possible for DWP to encrypt the NINo on the WPLS using the same encryption software to obtain an encrypted NINo. Using tax and NI activity, it should then be possible to arrive at a good estimate of how long migrants stay in the country. This would be invaluable since the MWS data does not provide any information on length of stay.