

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 NSMAC13 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 Goldstein 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 NSMAC13 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 been 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 |