Government Statistical Service

Minutes of the 31st Meeting of the GSS Methodology Advisory Committee

24th May 2016

Drummond Gate, London

Hosted by the Office for National Statistics

Government Statistical Service

Contents

1.0 List of Attendees

Committee members present

Martin Axelson	Statistics Sweden
David Best	Office for National Statistics
Pete Brodie	Office for National Statistics
Siobhan Carey	Department for Business, Innovation and Skills
Tricia Dodd (Chair)	Office for National Statistics
David Firth	Warwick University
Brian Francis	Lancaster University
Denise Lievesley	Green Templeton College, Oxford
Denise Osborn	University of Manchester
John Pullinger	UK Statistics Authority
Jeff Ralph	Office for National Statistics
Peter Smith	University of Southampton
Patrick Sturgis	University of Southampton

Presenters

Robert Breton	Office for National Statistics
Adam Douglas	Office for National Statistics
Jane Naylor	Office for National Statistics
Sebnem Oguz	Office for National Statistics
Paul Smith	University of Southampton
Lucy Vickers	Office for National Statistics

Apologies

James Gillan	Northern Ireland Statistics and Research Agency
Jouni Kuha	London School of Economics
Bill Oates	Office for National Statistics
Heather Savory	Office for National Statistics
Adrian Smith	UK Statistics Authority
Richard Smith	University of Cambridge

Others present

Sarah Adams	Office for National Statistics
Katie Connolly	Office for National Statistics
Marie Cruddas	Office for National Statistics
Silvia Manclossi	Welsh Government
Chris McGowan	Welsh Government
Lawrence Powell	Office for National Statistics
Emma Timm	Office for National Statistics
Bella Wheeler (Secretary)	Office for National Statistics

2.0 Administration

Tricia Dodd (chair) welcomed everyone to the 31st GSS Methodology Advisory Committee meeting, (including new members Siobhan Carey, Denise Lievesley and David Best, ONS Director, attending for the first time) and gave apologies for those who could not attend the meeting. Attendees introduced themselves, briefly providing their background for the benefit of new members.

This is the last meeting for the secretary Bella Wheeler, Tricia introduced the new secretary Emma Timm and thanked Bella for her work in organising these meetings, and went on to inform the group that this is her last meeting as chair, as she is leaving ONS in July.

John Pullinger gave his personal thanks to Tricia for her contribution to methodology for the UK and for ONS in particular has been profound and certainly in the last two years (while he has been National Statistician) no one has been more active than Tricia in stepping into the very difficult spaces and when we have had requests from other government departments where they're trying to tackle some really complex and sensitive questions, Tricia has always been the person he has turned to when someone asks him a hard statistical question, so it has been great to have Tricia here, on behalf of all of us (ONS) thanked her for everything she has done throughout her career.

2.1 ONS and GSS news

Tricia provided the following ONS and wider GSS news;

- ONS welcomed the Bean review of economic statistics and are working through the implications of the recommendations of it.
- The 21st GSS Methodology Symposium taking place on 6th July at Department for Business Innovation and Skills, will be looking back over our 21 years and more importantly looking forward at what's to come in the future, John Pullinger will be giving one of the key note speeches.
- National Statistic Quality Review of the Living Costs and Food Survey was published on Friday 20th May.
- Quality conference being held in Madrid 1st-3rd June – a good selection of ONS and UKSA papers will be presented.

Paul Smith added that the International Conference on Establishment Surveys (ICES-V) is taking place in Geneva from 20th-23rd June.

2.2 Methodology news

Tricia informed the committee that (on the back of the Bean Review) an independent review of ONS' methodology functions by Andrew Garrett (a senior methodologist currently working with the Royal Statistical Society (RSS) as their Data Science Ambassador) is now underway. It will look at how ONS address methodology both now and more importantly the kind of services ONS are going to need in the future, and will be reported back to ONS at the end of July.

2.3 Minutes and progress from GSS MAC 30

Tricia referred to the progress since GSS MAC 30 noted in the meeting booklet, and minor edits to the minutes of the 30th GSS MAC meeting were requested and agreed.

Government Statistical Service

3.0 Papers presented

3.1 Paper 1: Developing an approach to addressing potential discontinuities in the new National Survey for Wales

Authors	Paul A. Smith[1], Jan van den Brakel[2], Nikos Tzavidis[1]	[1]University of Southampton [2]Maastricht University
Presented by	Paul Smith	University of Southampton
Discussant	David Firth	Warwick University

Presentation

Paul Smith identified that the University of Southampton has been working closely with Welsh Government to develop an approach to address potential discontinuities in the new National Survey for Wales (NSW). Paul noted that the presentation was about the proposed approach but that the work had not been done yet.

Paul began by detailing the differences between the new NSW and the five surveys that it would be replacing. The new design is unclustered over the year, similar to the design of the Labour Force Survey in that the person-level survey may be regarded (mainly) as a one-stage cluster sample of people, with the clusters (or primary sampling units) being the households, with some sub-sample questions. Paul explained that there was a pilot survey that overlapped with the old survey, which will be used to investigate the differences between this and the new design.

Paul then outlined the assumptions that will need to be made when assessing discontinuities. The pilot will have different variances due to clustering and smaller sample sizes, but the same measurement errors. Timing differences of pilot and last periods of existing surveys can be ignored.

Paul proposed a small area model approach (van den Brakel et al) for measuring discontinuities. Small areas would include age by sex groups at national level. Two options were proposed:

1. Estimate $\hat{\theta}_i$ and $\hat{\theta}_i^*$ and take the difference
2. Estimate $\hat{\theta}_i - \hat{\theta}_i^*$ directly

It was suggested that the second option may be easier but it was questioned if there was a good predictor for this discontinuity. A disadvantage of the first option was that there would be a design based variance for $\hat{\theta}_i^*$ and a model based variance for $\hat{\theta}_i$.

Paul explained the definition of a discontinuity in this context but that this difference may be important but not significant due to the variances and questioned how to explain this to users of the data. A possible way of testing the differences is by using multiple t-tests but this would lead to lower power. Benjamini-Hochberg method helps with this.

Paul described the options for adjusting the series to account for discontinuities. A factor could be added to the old series to bring it up to the level of the new estimates or subtracted from the new estimates to bring it down to the level of the estimates in the old series. There was also the option of multiplying proposed.

Paul indicated that in the future (when 2-3 years of data has been collected) it may be possible to use time series methods to adjust for discontinuities. This will borrow strength over time not areas.

The main questions posed in the paper were:

Question 1: What is the best way to combine the model variance and sampling variance? How would you best interpret a statistic which combines a sampling and a modelling process?

Question 2: Does the MAC have any guidance on how to present and interpret estimates of discontinuities when the power to detect differences from zero is low?

Question 3: How is it best to present the results of the Benjamini-Hochberg adjustment procedure? There doesn't seem to be an approach which produces confidence intervals adjusted for multiple testing using the B-H procedure.

Question 4: What is the best approach to make an adjustment (or not) for an important discontinuity with a large standard error?

Discussant response

David Firth suggested that not all the steps proposed need to be done and that hypothesis testing may not be the best idea.

David suggested that the separate surveys should not be analysed separately but that a Bayesian model should be used to model them together.

David noted that the use of thresholds can be a bad idea even if well determined. It may result in inherent discontinuity in what is done therefore leading to multiple discontinuities causing interpretation to be more confusing. David suggested that reporting that there is a discontinuity may be a better approach. David then detailed the methodology of the Bayesian approach.

Some advantages of the Bayesian approach are:

- Assumptions are explicit
- There is structure in the model for the differences providing us with smaller standard errors
- No ad-hoc combinations of variances
- Clear to present results

David noted his scepticism behind the reasoning why the Bayesian approach had not been used in the van den Brakel paper.

In response to the questions asked in the paper David suggested that the Bayesian approach would help, presenting results in a coherent way making them easier to interpret.

Open discussion

Patrick Sturgis suggested that a causal approach may be better. There are potential changes due to non-response, measurement differences and general differences in the world. These are all potentially confounding and could negate each other. As a result we may not see genuine changes in the population due to the other changes.

Brian Francis raised concerns about the scale of the model. The sample size is quite small for some of the small areas and therefore some of the proportions could be quite close to 0 or 1. It may be worth modelling log odds rather than differences in proportions.

Paul responded to Brian by explaining that a linear model is preferred as it is easier to interpret but they will be careful if the proportions are close to 0 or 1.

Denise Osborn agreed that the Bayesian approach is a good idea. Denise raised concerns about the threshold that's been set. A difference of 0.05 may be big in some cases and small in others. If the numbers are being used to determine funding for certain projects this may cause issues.

John Pullinger stated that the user needs should be an important consideration. John suggested that it would be nice to get a general method for all to use. Work conducted by ONS on mode effects e.g. for the Volunteering survey and National Accounts should be considered. They may be able to provide advice when adjusting the series.

Action 3.1 – ONS (Tricia to delegate appropriately) to share work on mode effects with Paul and Welsh Government, and provide advice on adjustments if requested.

David Firth agreed with Patrick Sturgis on the idea of using a causal approach but suggested that measurement differences and time differences would be visible as separate things. The model over time is normally smooth. In response to Brian Francis, David agreed that a linear model would generally be preferred over the log odds model and gave the elections model as an example.

Patrick suggested that there may be practical difficulties getting a smooth time series for the old data due to the sparseness of some of the surveys and re-iterated that a warning on variables should be produced for the old and new estimates.

Martin Axelson noted that the pilot would provide a model based estimator but the full sample would provide a GREG estimator and posed the question: Are they both estimating the same thing?

Paul indicated that a design based estimate would be used initially then model based estimation will be used for the small area estimates. A model based adjustment would then be applied to the design based estimate. Paul also noted that it could all be model based with more data but this will need to be kept up to date.

It was noted that the threshold of 0.05 was set, not necessarily for all variables but to allow for a simplified response for the wide range of stakeholders.

3.2 Paper 2: 2021 Census Coverage Adjustment Methodology

Authors	Sebnem Oguz and Owen Abbott	Office for National Statistics
Presented by	Sebnem Oguz	Office for National Statistics
Discussant	Brian Francis	Lancaster University

Presentation

Sebnem Oguz explained the role of the Census Coverage Survey in determining those that have been missed from the Census and the need to adjust based on this, and made reference to the 2021 Census paper presented at the 30th MAC meeting. In a traditional survey, weights are created to represent the whole population, but for the Census, users do not want weights, instead a complete microdataset is desired, so imputation is needed to account for the two types of non-response (missing individuals and missing whole households).

Sebnem detailed the method used in the 2011 Census. First missing gaps in households are filled then extra households are created to adjust the dataset to make it representative of the population. Problems with this issue are detailed in the paper including convergence issues. The question was posed: Do we really need these two steps?

Sebnem detailed a new operational research method that could be used as an alternative called Combinatorial Optimisation (CO). An example of how the method works was also provided. This method takes donor households from the same area to make up for the cases that we know are missing given constraints that are provided. This method works better with more constraints. This is an iterative method that is repeated until the estimate is close enough.

Sebnem informed the MAC that this process has been applied in an experiment containing five areas that were selected as representative from the last Census and the method was compared with the previous one. As a result the following were found:

- New method was much faster but this could have been due to different programming languages being used
- New method was easy to test multiple times to see if there were differences for each run
- Results were much closer when using the new method

The main questions posed in the paper were:

Question 1: Should we pursue research into the use of Combinatorial Optimisation for un-benchmarked variables?

Question 2: Does the committee agree that a Combinatorial Optimisation approach is worth pursuing?

Question 3: Does the committee have any comments on the suggested alternative approaches for adjusting the Census database?

Question 4: Can the committee suggest any other alternative approaches?

Question 5: Does the committee have any further suggestions for improving the Coverage adjustment methodology?

Discussant response

Brian Francis began with a disclaimer that he is not an expert in the area so his role would be to feedback on what he understood the problem to be and raise some issues.

Brian referred to the paper from the previous MAC meeting about Dual-system Estimation. The constraints are not absolute constraints.

Brain expressed concerns about the re-use of cases to impute for those that are missing. This was possible. Is this possible in CO? Concerns were also raised about the possibility of multiple starting points. Will this lead to different solutions? There was interest expressed in the stability of variables not related to those used in the imputation process over multiple runs. He also questioned what would happen at the lower (eg SOA) level? He wasn't sure that this method will work for geographic breakdowns.

Brian suggested fitting a log linear model if the constraints aren't necessarily exact and the use of administrative data where appropriate.

In response to the questions from the paper Brian said yes to questions 1 and 2 before suggesting a model based approach in response to question 3.

Open Discussion

Peter Smith posed the question: If through imputation, the database fails to meet the constraint, which number should be given as the population total?

David Best asked if the Address Register and administrative data sources could be incorporated into these estimates and it was agreed that this was a possibility.

John Pullinger mentioned the GSS Methodology symposium's theme for this year is innovation and expressed this work as being a case of innovation that he was very impressed with. He suggested that the methodology could be generalised so that a similar approach could be used when considering administrative data. John cautioned that using this method it could be possible to select an anomalous household as a donor but that the weights given to the donor households could help with this – he noted that the highest weight given to a case was 19, which is really impressive.

Sebnem clarified that they were considering the imputation methods not the estimation method used to construct the constraints.

John said that the geography question is an important one. **Marie Cruddas** said that the team do small area estimation (SAE) for local authorities and have some idea of quality at that level, but not below, they rely on 94% coverage, but need to incorporate uncertainty. She said there are important benefits to doing this work quickly – **David Best** said he would make a note to ensure a PC is available to run Fortran on for the Census team.

Martin Axelson suggested raking as another method in response to question 3 posed by the paper – calibrate on estimated totals at individual and household level simultaneously, impute for the integer part, calibrate toward same totals and repeat – might still be left with small problems.

Paul Smith stated that this concept works (at least the first step). **Sebnem** said they had tried raking and found it worked for household variables, but not for individual ones, and stressed that imputation was the issue.

Martin added another idea would be to put caps on the values of the weights.

David Firth explained that this is an optimisation problem and suggested that the number of times a household can be selected should be included in the constraints. Looking at households from other areas and creating synthetic households may also help. It may be the case that the households that are not captured are different from all of the households that are captured. Pete Brodie agreed with this.

Paul explained that care would need to be taken in the construction of synthetic households so that unrealistic households are not constructed (such as a household with two parents in their 80s and a child aged 2). Marie said that with the two forms of missingness can pick from those that have responded, and take out replicates.

Jeff Ralph suggested that the standard optimisation problem highlighted in the report could be solved by setting parameters to avoid local minima.

4.0 Project updates

4.1 Plans for use of administrative data

Presented by	Adam Douglas and Lucy Vickers	Office for National Statistics
Discussant	Martin Axelson	Statistics Sweden

Presentation

Adam Douglas and Lucy Vickers (head of ONS Admin Data Division) presented an overview of ONS' ambitious plan to maximise key data sources - where there is great potential (and challenges) –saying they were looking for a steer on where to target their efforts.

Administrative data was defined as a by product of large scale administrative systems like health or crime which over time have started to be used internally mainly for monitoring or legislative purposes. For example HMRC link various types of data, such as benefit claimants with tax payments.

There are potential uses of this data across the whole of the GSS and have been brought to the forefront by things such as the Bean Review, the Queen's recent speech and a push from Parliament. Sweden already use a lot of administrative data for their statistics and other countries are now looking at this more and more.

Adam explained that the first consideration is to understand the concepts used in the data. Contextually admin data is not being collected for statistical purposes and therefore differs from what we are looking for so these differences need to be understood in order to be able to manipulate the data for these purposes. Admin data can be thought of as a sampling frame, you need to understand it in that way before you can be able to match it to your needs. There is also an issue around the completeness of the data. Each data source will be unique in its structure, uses and understanding so understanding the data itself is an important start to using it effectively.

Once you have the data source defined it is a matter of assimilating the data into the current information (whether survey or other statistical work). Although this could be a matter of simply replacing survey data, there are other uses such as supplementing the data or even helping to improve the sampling methods which could be considered. The Australian Bureau of Statistics use revenue and customs data for example to replace small business data in their Annual Business Survey while still sending forms to medium and large businesses.

Admin data can also be used to fill in coverage gaps in areas which survey data is known to undercover, such as possibly using HMRC data to supplement the Annual Survey of Hours & Earnings which has difficulty sampling the wealthy end of the pay range.

The data could also be used as a tool to develop a sampling frame (such as identifying individuals with required characteristics to survey) or as a way of evaluating a survey's effectiveness, improving the survey quality rather than directly replacing it.

There are however some issues with using administrative data such as how to use it for forecasting, back casting or now-casting and linking to historical data.

Linking across admin data sources also has huge potential and may give a broader knowledge about the information, for example Statistics New Zealand use tax data as well as their death register to help identify migration out of the country (individuals who have had no economic activity but have not been registered as deceased are identified as emigrants).

Administrative data could also help reduce non-response and sampling bias, for example ONS business surveys don't all currently collect information on sole traders but the gaps could be supplemented with administrative data, similarly missingness due to non-responses could also be overcome.

There are also technical challenges for the data including getting hold of, storing and investigating these very large datasets.

Future data strategies need to incorporate administrative data into the plans and understand their potential uses.

The main questions posed were:

Question 1: What do the committee consider to be the biggest technical challenges to methodology before the data are actually ingested into ONS systems? Ie those technical challenges before the data are 'used in anger' within ONS?

Question 2: What factors do the committee see as the biggest technical challenges to integrating the data with ONS surveys?

Question 3: Much of the wider business strategy focuses on replacing survey with admin data. This will not always be possible (in the short to medium-term) so what other uses/benefits (technically) can the committee see arising from admin data?

Question 4: Which specific economic and social survey outputs do the committee think admin data will be (technically) most beneficial for: and least beneficial for?

Question 5: Which do the committee feel will be the biggest technical challenge; forecasting using admin data as a correlate (where appropriate) or integrating admin data into historic survey series/back-casting?

Discussant response

Martin Axelson addressed the discussion points and pointed out that admin data was similar to Big Data but had a few distinct differences:

- Volume may be large, but manageable as it is already in a database
- Velocity is high, but under control
- Variety between sources, but it is structured
- Veracity will vary between sources, dependent upon use

It is a major technical challenge to receive (or retrieve), structure and store the data. Sweden use Tratten to receive 10,000 files a year and this is helped by having a good framework for the data and the infrastructure & policies in place to deal with the data effectively. There is a prerequisite to understand the drivers behind the dataset such as why and how it is being collected.

Control is also a problem as you do not control changes in that dataset so it is important to have good communication where possible to ensure any changes are noted and understood.

The challenge is to transform data from being administrative to statistical - this will depend on the intended purpose/ use of the admin data, for example if it is being used as a sampling frame or register it would need different manipulation than if it were to be used as a statistical output/product itself.

There has to be some element of compromising user needs and the uses of this data as you may find different users have different demands.

It requires architecture in terms of organisation, infrastructure and standard procedures etc.

It should be remembered that there are many ways to use the data within the production process, not just replacing surveys. Improving survey quality is an important use and shouldn't be forgotten.

Modelling is required especially for back casting and forecasting as well as linking to historical data, as continuity in data can be a problem.

Open Discussion

Peter Smith noted there are some methodological issues which weren't covered in the presentation but noted the importance of understanding what the data represents.

Brian Francis said that the issues of data privacy and consent of users had not been mentioned, adding that public perception and trust in government are big issues.

Adam said the privacy and consent issues are being looked at but it was felt that the public would be more trusting of ONS linking other administrative data sources than other Government departments linking them themselves and that ONS would limit what they do with administrative data to less than what the home department are doing (for the data's original purpose). **David Best** added that we are only legally allowed to use administrative data for statistical purposes.

Brian also noted that Sweden has a population register where as in the UK people are a lot more concerned about giving out data and linking records.

Patrick Sturgis mentioned that from the academia side of things, getting admin data seems to be a very slow process and very complicated due to privacy and legal issues but felt that ONS might have less of these constraints or at least that is the public perception – if you asked public for permission to link their data, might find they think we are already linking it.

Lucy said that the constraints (legal and ethical) are still there for ONS and are being considered. The team is trying to create a portfolio approach of using the data so it is not just to replace surveys but also other uses. **Adam** mentioned the Admin Research Centre is working well with HMRC to produce a data pool (of anonymised data) but it is a slow process to receive updates.

Pete Brodie stated that ONS are at the beginning of this journey and there are lots of steps and statistical inferences to consider, such as:

- Imputation (missing data, no contact with respondents)
- Disclosure and privacy issues
- Metadata and changes to collection methods (with surveys we know every part of the process)
- Timeliness of data ('current' VAT data could be up to a year out of date)
- Combining admin data with historic and current datasets – possibly use Bayesian methods like in Paper 1

Pete is working closely with the Admin Data team to flesh out some of these issues but the hardest part is understanding the data source. We shouldn't be trying to do things the way we always have, need to start thinking of different approaches.

Martin Axelson pointed out how large the Admin Data Source Catalogue ONS produces currently is, and mentioned that although Stats Sweden have very good registers, covering everyone they're interested in, they are not used (as microdata) as often as you might think, instead they are used more as a sampling frame than as a data source (and thinks this is the same for the other Nordic states).

Peter Smith mentioned that with admin data there is no overall framework in which to make inferences from so this also needs to be thought about, along with measurement models (surveys).

David Best pointed out that surveys shouldn't be thought of as an all or nothing approach and we need to think in a radical way about what we want, the best way to gather the variables we need and dissemination, rather than assuming we are going to replace all surveys with administrative data.

Siobhan Carey said Pete's list was fairly comprehensive, and that you could look back at historic admin data but not get much from it. It would be useful to try and focus efforts on the Big Wins going forward i.e. concentrating on the core data we are interested in and how we can develop inference from that or whether it can be used as a verification of survey data to improve quality. She also mentioned learning lessons from other organisations that have combined data sources (such as Department of Education, BIS & HMRC) and noted that a lot of services are now going 'digital'.

The **Admin Data team** hope to bring some examples of what they are working on and more technical style papers to MAC in future.

Government Statistical Service

4.2 Update on Big Data and Classifying web scraped data using machine learning

Presented by	Jane Naylor and Robert Breton	Office for National Statistics
Discussant	Denise Lievesley	Green Templeton College

Presentation

Jane Naylor gave an overview of the ONS Big Data team, detailing their goals and approach. The main emphasis for the team is recognising the growth of big data sources and to develop data science skills. She also stressed the differences between big data and admin data.

Jane went on to talk about the different groups they are working with and various pilot projects being carried out. The benefits of these projects included: improved quality of data, reducing survey costs and more frequent data.

Jane talked about challenges relating to data quality, raising concerns about a lack of control over how big data is collected and classification issues. She then highlighted the issue of bias and stability, using the examples of age bias in Twitter users and comparing Google Trends data to Labour Force Survey data, respectively. She concluded these sources can be carefully used as early indicators.

Robert Breton took over to present web scraping and machine learning. He gave a background to web scraped data, which is 6500 daily quotes, taken from 3 supermarkets in 35 CPI item categories. Rob briefly explained the process of web scraping which led to the challenge of misclassification. The key word filtering solution shown was time consuming and instead machine learning was proposed by Rob as a solution.

Training data, classified by humans, is fed into a machine learning algorithm and then classified to a predicted class. The performance of the model can be judged by a confusion matrix on unseen data, where errors in prediction can be compared to actual results. The classifier is then integrated into the web scraping system, to classify as you stream and thus rapidly develop classifiers.

Rob concluded that supervised machine learning has sped up classification and that there is scope to apply this to other fields within ONS. He also stressed that there is a need for more training data, utilisation of supporting unsupervised methods and for greater use of grid search to optimise parameters.

The main questions posed were:

Question 1: What do the committee consider to be the key methodological challenges associated with the use of big data sources, tools and technologies for official statistics?

Question 2: Should ONS be focussing on developing in-house skills (such as machine learning) for cleaning, classifying and transforming new forms of data?

Question 3: Should ONS look to purchase cleaned data from third parties: essentially contracting out the cleaning, classifying and transforming new forms of data?

Question 4: Some machine learning methods are highly effective at classification, however, the data transformation and optimisation methods are hard to understand, explain and thus make transparent. What should be the level of trade-off between complexity and transparency?

Question 5: We are investigating the use of machine learning to assist in product classification; are there other approaches that you would recommend in classification tasks on high frequency data?

Discussant response

Denise Lievesley began by stressing the importance of big data sources within national statistical offices and was pleased with the level of collaboration by the Big Data team. She drew parallels with big data research being undertaken at Oxford Internet Institute. From a management point of view, she highlighted a need to plan the systems needed for non-statisticians and a recruitment initiative for multi-lingual experts. She also drew synergies with data scientists and computational biologists, as these are relatively new professions that have increasing demand. She added by saying that the American Statistical Association highlighted data science as the field with most potential.

Denise then went on to talk about what Jane and Robert did not cover. She mentioned that, instead of sample error, the largest risk of big data lies in its bias, relating to the Twitter data example. She stressed that it is important to engage this big data with current statistics for inference. Other issues not covered include correlation and causation. This relates to the example of Google Trends and Labour Force Survey data for immigration from Poland. She also raised the question on how we communicate uncertainty in these statistics and how they can be reproduced.

Denise mentioned the need for a lot of investment in big data. She used the example of switching from pen and paper to computer methods requiring a large initial investment, which could bring about management challenges.

She concluded that whilst there was good progress in this area, there is still a need for establishing the ethics, trust and transparency surrounding big data sources. As this is a rapidly growing area, it is essential that the public, politicians and advisors for new methods are carried on this journey.

Open Discussion

Siobhan Carey: Highlighted that the main risk is outside development, similar to their work. She went on to say that it can be done by aligning themselves with scientific disciplines and research councils. She said that this was important because there is no need to reinvent the wheel, but this is not BIS's natural space.

Brian Francis: Urged the big data team to keep on questioning their data sources. Relating to the age bias issue with Twitter data mentioned earlier, he reminded them that social trends are always changing. He used the example of the age demographic of Facebook users slowly getting older, and the younger generation using new forms of social media, such as Instagram. He also added that it is important for machine learning to incorporate human intervention.

Robert Breton: Agreed and stressed that the Big Data team are very critical of their data sources and they already keep the human intervention element in machine learning.

David Firth: Brought up the issue of statisticians not being equipped with a computer scientist's skill-set and vice-versa.

Denise Lievesley: Concluded by expressing the need for synergy within this project, relating to her point about computational biologists earlier.

5.0 AOB

David Best took the opportunity to talk to MAC members about ONS' Data Science Campus, which will be a home to improve capability in data science and statistics as well as running jointly funded research and focused projects for other government departments. Areas of focus under capability (entry level will be degrees and up), data science and engineering, will be:

- Urban space (social activities)
- Human Area (wellbeing, unemployment)
- Sustainability measures
- UK in a global context (flows)
- Digital Economy

Project proposals in these areas will be called for soon. David Johnson (formerly worked for Google) will be leading in the start up of the centre, with Heather Savory and John Pullinger sponsoring.

Tricia then concluded the meeting by summarising the topics covered during the day, thanking the committee for attending today and during her time as committee chair.

6.0 Summary of actions

Action 3.1 – ONS (Tricia to delegate appropriately) to share work on mode effects with Paul Smith and Welsh Government, and provide advice on adjustments if requested.

Secretary to inform members of next meeting date.