Australian Bureau of Statistics

1351.0.55.056

Research Paper

A Statistical Framework for Analysing Big Data

www.abs.gov.au

New Issue

Research Paper

A Statistical Framework for Analysing Big Data

Dr Siu-Ming Tam

Chief Methodologist
Australian Bureau of Statistics

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) TUE 30 JUN 2015

ABS Catalogue no. 1351.0.55.056

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics.
Where quoted, they should be attributed clearly to the author(s).

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper. For further information, please contact Dr Siu-Ming Tam, Chief Methodologist, Methodology and Data Management Division, on Canberra (02) 6252 6307 or email <Siu-Ming.Tam@abs.gov.au>.

EXECUTIVE SUMMARY

In this paper, it is contended that the threshold challenges that must be adequately addressed before Big Data sources can be used for the production of official statistics are the business case, the validity of statistical inference and data ownership and access issues.

The business case comprises business needs and benefits, and data ownership and access issues are particularly important where, as is commonly the case, the National Statistical Office is not the custodian of the Big Data source. Above all, given the expected inferential biases from Big Data – due to under-coverage, self-selection, missing values etc. – statistical methods must be developed before Big Data sources can be harnessed for the production of official statistics.

Using a Bayesian framework, this paper outlines necessary conditions – in particular, the Missing At Random condition – for valid statistical inference to be made for estimating or predicting finite population parameters (e.g. totals of population units), or for estimating the super-population parameters of statistical models (e.g. the regression coefficients of a linear regression model).

By assuming that Missing At Random conditions are fulfilled, the paper also provides an illustrative theoretical method for utilising satellite imagery data to predict crop areas and crop yields. The analysis assumes that the data are described by a dynamic logistic model for crop types and a dynamic linear model for crop yields. The method relies on using "ground truth" data from a random sample to calibrate the satellite imagery, and using the latter as covariates to predict the data of interest for the population not included in the random sample.

Finally, the paper outlines methods to address related statistical computing issues and proposes strategies for extending the model to provide a better fit to the observed data.

CONTENTS

A STATISTICAL FRAMEWORK FOR ANALYSING BIG DATA

Dr Siu-Ming Tam
Chief Methodologist
Australian Bureau of Statistics[1]

ABSTRACT

In this paper, it is contended that the threshold challenges that must be adequately addressed before Big Data sources can be used for the production of official statistics are the business case, the validity of statistical inference and data ownership and access issues.

Using statistical modelling, the paper outlines necessary conditions for addressing the biases inherent in Big Data sources when estimating parameters of a finite population or super-population model.

To illustrate the proposed statistical framework, the paper describes a method, based on State Space modelling, for utilising satellite imagery data to predict crop types and crop yields. The paper also outlines methods to address related statistical computing issues, and proposes strategies for extending the model to provide a better fit to the observed data.

[1] This paper was written in response to an invitation from the Editor of the *Survey Statistician* to provide some perspectives on the *American Association for Public Opinion Research (AAPOR) Task Force Report on Big Data* (Japec *et al.*, 2015).

1. INTRODUCTION

In a 2014 talk to the Victorian Branch of the Australian Statistical Society, Professor Terry Speed, an eminent mathematical statistician and winner of the 2014 Australian Prime Minister's Science Award, expressed surprise about the lack of visibility of statisticians in the Big Data debate, and said "…the absence of statisticians in Big Data activities is striking (to a statistician)". He also observed that there was generally lack of presence of statisticians in national and international conferences on Big Data.

In an article entitled "Big Data or Big Fail? The Good, the Bad and the Ugly and the Missing Role of Statistics", Iacus (2014) echoed Terry Speed's point about the role statistics and statisticians can play in the field of Big Data.

Against this background, I warmly welcome the well written and researched Report by the American Association for Public Opinion Research (AAPOR) Task Force (Japec *et al.*, 2015). The references provided in the Report would be very useful to statisticians who want to use Big Data or make a contribution to the Big Data debate.

I particularly like the report's comprehensiveness in raising the many different issues of Big Data, covering not only what it is and why it matters, but also the policy, technical and technology challenges facing users of Big Data in solving business problems or finding answers to societal questions.

As a practising official statistician, I find Section 7 of the AAPOR Report very interesting, and in particular, Sub-section 7.3 about combining Big Data and Survey Data. I would therefore devote most of my comments on this issue. I would also outline the preliminary work undertaken in the Australian Bureau of Statistics (ABS) to investigate into the business case and validity of harnessing certain Big Data sources for the regular production of official statistics.

2. THRESHOLD CHALLENGES FOR BIG DATA

Whilst the Report has outlined a number of key challenges for Big Data use and analysis, I would contend Business Case, using Big Data in statistically valid ways, i.e. Validity of Statistical Inference (page 22 of the Task Force Report) and Data Ownership (page 30) are the threshold challenges confronting official statisticians in the use of Big Data in the regular production of official statistics.

In saying this, I am not downplaying the other challenges such as Data Stewardship, Data Collection Authority, Privacy and Re-identification. National Statistical Offices (NSOs) are generally well set up and have developed capability to address these challenges. For example, many statistical offices have already developed methods, processes and procedures to address privacy and confidentiality issues in their statistical releases – see, for example, the Special Issue of the *Statistical Journal of the International Association of Official Statistics* on “Official Statistics and Micro Data: Access and Confidentiality” released in 2009 – which may be adapted to address releases based on, or supplemented by, Big Data. A detailed discussion of the Big Data challenges faced by NSOs are provided in Tam and Clarke (2015a). My contention is that if the threshold challenges cannot be overcome, i.e. there is no business case for using a particular Big Data source, if the Big Data source cannot provide valid statistical inferences, and if the Big Data source is not available to official statisticians, there is no question of using the Big Data source in regular statistical production, and the other challenges do not arise.

3. BUSINESS CASE

What is the Business Case of Big Data? Business case comprises business need – what business problems we want to solve and can Big Data be part of the solution – and business benefit – whether the benefit of Big Data as a solution does outweigh the costs?

Being a collective term for a diverse range of data sources (page 5), the business case for Big Data does vary from source to source. For example, there is clearly a business case in the use of Administrative Data (page 9) by official statisticians in the production of official statistics, e.g. in the use of birth, death and migration records to complement the data from population censuses to provide contemporary population estimates. Cargo manifests are used to produce trade statistics. Without these sources, it will not be possible to provide population estimates or trade statistics. In other words, these sources provide valuable information to fill a data gap.

However, I have heard of propositions such as "… let's bring all the Big Data into our organisation and then figure out what we want to do with it. And to effectively do this, let's upgrade our computer hardware, or software, because Big Data requires big data processing capabilities …". These propositions worry me as they put the cart (Big Data) before the horse (business problems) and treat "Big Data as a solution in search of a problem".

In my view, Big Data should only be used if it can:

- improve the product offerings of statistical offices e.g. more frequent release of official statistics, more detailed statistics, more statistics for small population groups or areas, or filling an important data gap – business need; or
- improve the cost efficiency in the production of official statistics – business benefit.

The AAPOR Report rightly points out (page 15) that the "costs and risks of realising these (i.e. Big Data) benefits are non-trivial". For example, in the case of satellite data, whilst the risk of not having access to the data is small given that most of these are available free of charge on the internet, the cost associated with creating the ground truth data and marrying them up with satellite data, at the observation unit e.g. a statistical local area the cost of storing, cleaning, processing, quality assuring and software development are substantial. In the case of the Australian Bureau of Statistics (ABS), while the business need for using satellite data, instead of direct data collection, to estimate crop areas and crop yields has been well established, the business benefit has yet to be assessed.

4. A POSSIBLE APPROACH TO USING BIG DATA FOR OFFICIAL STATISTICS

An approach which has recently been actively pursued at the ABS (Tam and Clarke, 2015b) for the use of Satellite data in official statistics production is to consider the $\mathrm{N} \times 1$ vector of measurements, $\mathbf{Y}_t$, of interest to the official statistician, e.g. crop areas or yields, at time t as a realisation of a super-population model, with the Big Data augmented with non-Big Data sources, $\mathbf{Z}_t$, treated as a (design) matrix of covariates for the model, i.e.

$$\mathrm{Y}_t = \mathrm{Z}_t \boldsymbol{\beta}_t + \mathbf{e}_t \tag{1}$$

and allowing the vector of regression parameters, $\boldsymbol{\beta}_t$, to change over time, i.e.

$$\boldsymbol{\beta}_t = \mathrm{H}_t \boldsymbol{\beta}_{t-1} + \boldsymbol{\varepsilon}_t \ . \tag{2}$$

Here N denotes the size of the finite population e.g. total number of land parcels. Equations (1) and (2) form the well-known State Space Model. Under this formulation, we consider that a sample, s_t, of units is chosen, e.g. a sample of observation units at time t, on which observations of the value of $\mathbf{Y}_{ot}$, where ‘o’ denotes observed (or responding) units, are obtained. Denote by ‘m’ the units in s_t on which there is no observation, i.e. missing data, and ‘r’, the units of s_t not selected in the sample, then the vector $\mathbf{Y}_t$ can be partitioned as $\mathbf{Y}_t = \left(\mathbf{Y}_{ot}, \mathbf{Y}_{mt}, \mathbf{Y}_{rt}\right)'$. State Space Models were used in Tam (1987) for predicting finite population parameters in finite population sampling.

Assuming that we can match these observed units to the corresponding units in the Big Data source and non-Big Data sources available to the statistician e.g. geographic location (in a survey, the linkage is automatic through the questionnaire as a collection instrument), and as can be seen from diagram 5.1 below, for every unit in the sample, s_t, one of the following two conditions will apply, namely, that there is a corresponding set of data from Big Data for the unit, and there is not. Denote by ‘B’ those units that have Big Data information, and ‘$\tilde{\mathrm{B}}$’ those that don’t. Then (1) can be re-written as:

$$\begin{bmatrix} \mathrm{Y}_{o_B t} \\ \mathrm{Y}_{m_B t} \\ \mathrm{Y}_{r_B t} \\ \mathrm{Y}_{o_{\tilde{B}} t} \\ \mathrm{Y}_{m_{\tilde{B}} t} \\ \mathrm{Y}_{r_{\tilde{B}} t} \end{bmatrix} = \begin{bmatrix} \mathrm{Z}_{o_B t} \\ \mathrm{Z}_{m_B t} \\ \mathrm{Z}_{r_B t} \\ \mathrm{Z}_{o_{\tilde{B}} t} \\ \mathrm{Z}_{m_{\tilde{B}} t} \\ \mathrm{Z}_{r_{\tilde{B}} t} \end{bmatrix} \boldsymbol{\beta}_t + \begin{bmatrix} \mathbf{e}_{o_B t} \\ \mathbf{e}_{m_B t} \\ \mathbf{e}_{r_B t} \\ \mathbf{e}_{o_{\tilde{B}} t} \\ \mathbf{e}_{m_{\tilde{B}} t} \\ \mathbf{e}_{r_{\tilde{B}} t} \end{bmatrix} \tag{3}$$

Note that (3) can be extended to Generalised Linear Models and Generalised Linear Mixed Models – see the penultimate section of this paper.

Let $\mathbf{I}_t$, $\mathbf{R}_t$ and $\mathcal{R}_t$ denote random variables representing sampling, response and Big Data under-coverage processes respectively. These are column vectors whose i-th element is given by $\delta_i^{(I_t)}$, $\delta_i^{(R_t)}$ and $\delta_i^{(\mathcal{R}_t)}$ respectively, which is 'one' if the i-th unit is in the sample, responded or covered in the Big Data respectively; and 'zero' otherwise.

The inference problem under the model (2) and (3) can then be stated as follows:

1. The data for inference for the finite population, say the population total, $\mathbf{1}'\mathbf{Y}$, at time t are

$$\mathbf{D}^{(t)} = \left\{\mathbf{Y}_{o_B 1}, \mathbf{Y}_{o_{\tilde{B}} 1}, \mathbf{Z}_{o_B 1}, \mathbf{Z}_{m_B 1}, \mathbf{Z}_{r_B 1}, \ldots, \mathbf{Y}_{o_B t}, \mathbf{Y}_{o_{\tilde{B}} t}, \mathbf{Z}_{o_B t}, \mathbf{Z}_{m_B t}, \mathbf{Z}_{r_B t}\right\}$$

and $$\mathbf{P}^{(t)} = \mathbf{P}_1^{(t)} \cup \mathbf{P}_2^{(t)}$$

where $$P_1^{(t)} = \left\{\mathbf{I}_1, \mathbf{R}_1, \ldots, \mathbf{I}_t, \mathbf{R}_t\right\}$$

and $$P_2^{(t)} = \left\{\mathcal{R}_1, \ldots, \mathcal{R}_t\right\}.$$

2. Model-assisted methods (Särndal *et al.*, 1992) and model-based methods (Chambers and Clark, 2012), including Bayesian methods (Puza, 2013), may be applied for making inference.

3. Whatever method of inference is used, the official statistician needs to understand, or make assumptions, about the processes leading to the missing and non-sample data, i.e. how those highlighted in black in equation (3) come into being; Where missing at random conditions are not met (see Section 5 below), modelling for the missing and non-sample selection processes have to made. For Big Data sources, this can be very challenging, if not insurmountable.

5. VALIDITY OF STATISTICAL INFERENCES

I welcome the attempt by the Task Force to provide a total error framework for Big Data (page 18), and Couper (2013) provides a good description of the types of errors encountered in Big Data.

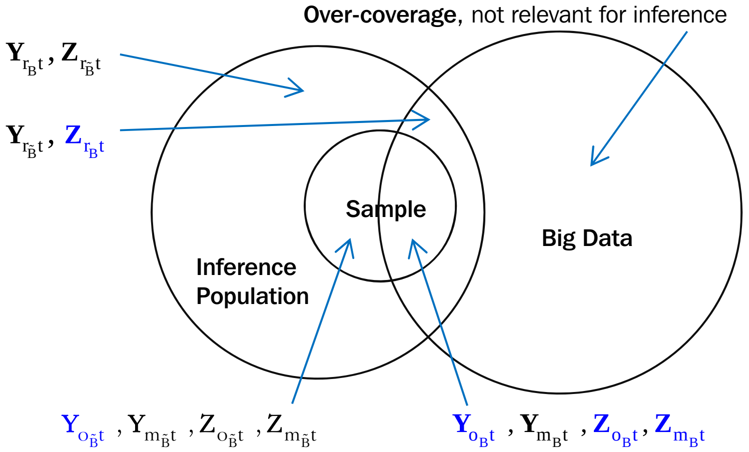
I cannot agree more strongly with the Report that "… using Big Data in statistically valid ways is challenging and one misconception is the belief that the volume of the data can compensate for any other deficiency in the data (Big Data Hubris)" (page 22). Unlike sampling errors, non-sampling errors will not be reduced by increasing the sample size. Likewise, correlation is not the same as causality. In a recent article in *Significance*, entitled "Big Data, Big Mistake?", Harford (2014) showed how such a misunderstanding can have fatal consequences. The Report's reference to Fan *et al.* (2014) is particularly valuable to those Big Data enthusiasts who believe that size is everything!

To explore the conditions for validity of statistical inference, we will depict the relationship between a particular Big Data source (e.g. satellite imagery data) and the target population of interest (e.g. the land parcels) to the official statistician, in diagram 5.1 below. As well, we will make the simplified (but not always true, e.g. social media data) assumption that the unit of interest in the target population will appear in the Big Data source, if at all, only once. This is to ensure the possibility of making an unique linkage between the $\mathbf{Y}$ value of a unit in the target population and the corresponding $\mathbf{Z}$ values from Big Data (and non-Big Data sources). (Note that if there are multiple appearances, an approach that may be adopted would be randomly choose one appearance where the appearances are homogeneous, include an additional covariate where there is structured heterogeneity, or use a repeated measures model (Denham *et al.*, 2011) where the appearances are sufficiently heterogeneous.)

The joint areas of the two big circles in diagram 5.1 are divided into three segments – under-coverage, i.e. information of interest to the official statistician but not available from Big Data; over-coverage, i.e. information available from Big Data that is of no interest; and finally, information of interest and available. Also, the 'system' can be described as comprising a data process, state process and censoring process, with prior distributions $f(\boldsymbol{\varphi})$, $f(\boldsymbol{\theta})$ and $f(\boldsymbol{\phi})$ with known hyper-parameters.

Under the approach advocated in this paper, I assume that a probability sample (so as to fulfil the non-informative sampling conditions for descriptive and analytic inferences – see (6) and (10) below) is drawn from the population of interest, from which observations are made. These, combined with the corresponding Big Data for the same observation units, are used to provide the posterior distribution of the model parameters – the Estimation step. The resultant posterior distribution, together with the Big Data for the non-sampled units, are then used to predict the values of these units using the predictive distribution – the Prediction step.

5.1 Integrating designed data with found data


Over-coverage, not relevant for inference
$Y_{r_{\tilde{B}}t}, Z_{r_{\tilde{B}}t}$
$Y_{r_{\tilde{B}}t}, Z_{r_{B}t}$
Sample
Big Data
Inference Population
$Y_{o_{\tilde{B}}t}, Y_{m_{\tilde{B}}t}, Z_{o_{\tilde{B}}t}, Z_{m_{\tilde{B}}t}$
$Y_{o_{B}t}, Y_{m_{B}t}, Z_{o_{B}t}, Z_{m_{B}t}$

Values of Y in units denoted by '$\tilde{B}$' not available due to under-coverage.

Conceptualise missing values due to Non-response Process, $\mathcal{R}_t$, applied to Big Data

Values of Y in units denoted by 'r' not available due to the sampling process, I_t.

Values of Y in units denoted by 'm' not available due to a missing process applied to the sampled data, R_t.

At time $= t$, State Space Process comprising:

Data Process – $f(Y_t;\varphi)$

State Process – $f(\beta_t,\theta)$
assumed to be Markovian.

'Censoring Processes' –
$f(I_t,\phi), f(R_t,\phi), f(\mathcal{R}_t,\phi)$

Parameter Models –
$f(\varphi), f(\theta), f(\phi)$

Data – $D^{(t)}, P^{(t)}$

Blue denotes observed/available.

5.1 Descriptive inferences

Under a Bayesian framework, the predictive inference of Y_t, $f\left(Y_t \mid D^{(t)}, P^{(t)}\right)$, given the data $D^{(t)}$ and $P^{(t)}$ – which I shall denote by $\left[Y_t \mid D^{(t)}, P^{(t)}\right]$ to simplify notation – is given by

$$\left[Y_t \mid D^{(t)}, P^{(t)}\right] = \frac{\left[P_1^{(t)} \mid Y_t, D^{(t)}, P_2^{(t)}\right]\left[Y_t, D^{(t)}, P_2^{(t)}\right]}{\left[P_1^{(t)} \mid D^{(t)}, P_2^{(t)}\right]\left[D^{(t)}, P_2^{(t)}\right]}$$
$$= \left[Y_t \mid D^{(t)}, P_2^{(t)}\right],$$

provided that
$$\left[P_1^{(t)} \mid Y_t, D^{(t)}, P_2^{(t)}\right] = \left[P_1^{(t)} \mid D^{(t)}, P_2^{(t)}\right]. \tag{4}$$

Assuming further that the finite population sampling and non-response processes at time τ_1 and τ_2 are independent for $\tau_1 \neq \tau_2$ and $\tau_1, \tau_2 = 1, \ldots, t$, sufficient conditions for (4) to hold are

$$\left[R_\tau \mid I_\tau, Y_\tau, D_\tau, \mathcal{R}_\tau\right] = \left[R_\tau \mid I_\tau, D_\tau, \mathcal{R}_\tau\right] \tag{5}$$

and
$$\left[I_\tau \mid Y_\tau, D_\tau, \mathcal{R}_\tau\right] = \left[I_\tau \mid D_\tau, \mathcal{R}_\tau\right] \tag{6}$$

for $\tau = 1, \ldots, t$.

Equation (6) holds for probability sampling, and Equation (5) holds if the non-response mechanism is missing at random (MAR) (Rubin, 1976). See, for example, Little and Rubin (2002) for response process modelling in which MAR does not hold.

Now,

$$\begin{aligned}\left[\mathrm{Y}_\mathrm{t}\,\middle|\,\mathrm{D}^{(\mathrm{t})},\mathrm{P}_2^{(\mathrm{t})}\right] &\propto \int\left[\mathrm{Y}_\mathrm{t},\mathrm{D}^{(\mathrm{t})},\mathrm{D}_\mathrm{c}^{(\mathrm{t})},\mathrm{P}_2^{(\mathrm{t})}\right] d\mathrm{D}_\mathrm{c}^{(\mathrm{t})}\\ &= \int\left[\mathrm{P}_2^{(\mathrm{t})}\,\middle|\,\mathrm{Y}_\mathrm{t},\mathrm{D}^{(\mathrm{t})},\mathrm{D}_\mathrm{c}^{(\mathrm{t})}\right]\left[\mathrm{Y}_\mathrm{t},\mathrm{D}^{(\mathrm{t})},\mathrm{D}_\mathrm{c}^{(\mathrm{t})}\right] d\mathrm{D}_\mathrm{c}^{(\mathrm{t})}\\ &= \int\left[\mathrm{Y}_\mathrm{t},\mathrm{D}^{(\mathrm{t})},\mathrm{D}_\mathrm{c}^{(\mathrm{t})}\right] d\mathrm{D}_\mathrm{c}^{(\mathrm{t})}\\ &\propto \left[\mathrm{Y}_\mathrm{t}\,\middle|\,\mathrm{D}^{(\mathrm{t})}\right]\end{aligned}$$

provided that $$\left[\mathrm{P}_2^{(\mathrm{t})}\,\middle|\,\mathrm{Y}_\mathrm{t},\mathrm{D}^{(\mathrm{t})},\mathrm{D}_\mathrm{c}^{(\mathrm{t})}\right] = \left[\mathrm{P}_2^{(\mathrm{t})}\,\middle|\,\mathrm{Y}_\mathrm{t},\mathrm{D}^{(\mathrm{t})}\right], \qquad (7)$$

where

$$\mathrm{D}_\mathrm{c}^{(\mathrm{t})} = \left\{\mathrm{Y}_{\mathrm{m_B}1},\mathrm{Y}_{\mathrm{r_B}1},\mathrm{Y}_{\mathrm{m}_{\tilde{\mathrm{B}}}1},\mathrm{Y}_{\mathrm{r}_{\tilde{\mathrm{B}}}1},\mathrm{Z}_{\mathrm{o}_{\tilde{\mathrm{B}}}1},\mathrm{Z}_{\mathrm{m}_{\tilde{\mathrm{B}}}1},\mathrm{Z}_{\mathrm{r}_{\tilde{\mathrm{B}}}1},\ldots,\mathrm{Y}_{\mathrm{m_B}\mathrm{t}},\mathrm{Y}_{\mathrm{r_B}\mathrm{t}},\mathrm{Y}_{\mathrm{m}_{\tilde{\mathrm{B}}}\mathrm{t}},\mathrm{Y}_{\mathrm{r}_{\tilde{\mathrm{B}}}\mathrm{t}},\mathrm{Z}_{\mathrm{o}_{\tilde{\mathrm{B}}}\mathrm{t}},\mathrm{Z}_{\mathrm{m}_{\tilde{\mathrm{B}}}\mathrm{t}},\mathrm{Z}_{\mathrm{r}_{\tilde{\mathrm{B}}}\mathrm{t}}\right\}$$

represents the set of unobserved response variables and covariates in (3) for time 1 to time t .

Assuming that the under-coverage 'processes' for Big Data at time τ_1 and τ_2 are independent for $\tau_1 \neq \tau_2$ and $\tau_1,\tau_2 = 1,\ldots,\mathrm{t}$, sufficient conditions for (7) to hold are:

$$\left[\mathcal{R}_\tau\,\middle|\,\mathrm{Y}_\tau,\mathrm{D}_\tau,\mathrm{D}_{\tau\mathrm{c}}\right] = \left[\mathrm{R}_\tau\,\middle|\,\mathrm{Y}_\tau,\mathrm{D}_\tau\right], \qquad (8)$$

where $$\mathrm{D}_\tau = \left\{\mathrm{Y}_{\mathrm{o_B}\tau},\mathrm{Y}_{\mathrm{o}_{\tilde{\mathrm{B}}}\tau},\mathrm{Z}_{\mathrm{o_B}\tau},\mathrm{Z}_{\mathrm{m_B}\tau1},\mathrm{Z}_{\mathrm{r_B}\tau}\right\}$$

and $$\mathrm{D}_{\tau\mathrm{c}} = \left\{\mathrm{Y}_{\mathrm{m_B}\tau\mathrm{c}},\mathrm{Y}_{\mathrm{r_B}\tau\mathrm{c}},\mathrm{Y}_{\mathrm{m}_{\tilde{\mathrm{B}}}\tau\mathrm{c}},\mathrm{Y}_{\mathrm{r}_{\tilde{\mathrm{B}}}\tau\mathrm{c}},\mathrm{Z}_{\mathrm{o}_{\tilde{\mathrm{B}}}\tau\mathrm{c}},\mathrm{Z}_{\mathrm{m}_{\tilde{\mathrm{B}}}\tau\mathrm{c}},\mathrm{Z}_{\mathrm{r}_{\tilde{\mathrm{B}}}\tau\mathrm{c}}\right\}.$$

Note that $\left[\mathcal{R}_\tau\,\middle|\,\mathrm{Y}_\tau,\mathrm{D}_\tau,\mathrm{D}_{\tau\mathrm{c}}\right] = \left[\mathrm{R}_\tau\,\middle|\,\mathrm{Y}_\tau,\mathrm{D}_\tau\right]$, for $\tau = 1,\ldots,\mathrm{t}$, may be satisfied for certain Big Data sources e.g. administrative data, but not others e.g. data from social media where participation is self-selected.

Where (4) and (7) are satisfied, $\left[\mathrm{Y}_\mathrm{t}\,\middle|\,\mathrm{D}^{(\mathrm{t})},\mathrm{P}^{(\mathrm{t})}\right] \propto \left[\mathrm{Y}_\mathrm{t}\,\middle|\,\mathrm{D}^{(\mathrm{t})}\right]$. In other words, the sampling, missing data and under-coverage processes can be ignored when making inference about Y_t .

Where (7) is not fulfilled, predictive inferences for Big Data will have to be based on $\left[Y_t \middle| D^{(t)}, P_2^{(t)} \right]$, which in turn requires modelling of $P_2^{(t)}$.

Prediction with missing covariates can be a very challenging problem. See, for example, Chapter 4 of Wu (2010) for possible methods and references to tackle this issue.

5.2 Analytic inferences

The posterior distribution of the parameters, θ and φ, is given by

$$\begin{aligned} \left[\theta, \varphi \middle| D^{(t)}, P^{(t)} \right] &= \int \left[\theta, \varphi, D_c^{(t)} \middle| D^{(t)}, P^{(t)} \right] dD_c^{(t)} \\ &\propto \int \left[P^{(t)} \middle| D^{(t)}, D_c^{(t)}, \theta, \varphi \right] \left[D^{(t)}, D_c^{(t)}, \theta, \varphi \right] dD_c^{(t)} \\ &\propto \left[\theta, \varphi \middle| D^{(t)} \right] \end{aligned}$$

provided that $$\left[P^{(t)} \middle| D^{(t)}, D_c^{(t)}, \theta, \varphi \right] = \left[P^{(t)} \middle| D^{(t)}, \theta, \varphi \right]. \tag{9}$$

Sufficient conditions for (9) to hold are

$$\left[P_1^{(t)} \middle| D^{(t)}, D_c^{(t)}, P_2^{(t)}, \theta, \varphi \right] = \left[P_1^{(t)} \middle| D^{(t)}, P_2^{(t)}, \theta, \varphi \right] \tag{10}$$

and $$\left[P_2^{(t)} \middle| D^{(t)}, D_c^{(t)}, \theta, \varphi \right] = \left[P_2^{(t)} \middle| D^{(t)}, \theta, \varphi \right]. \tag{11}$$

Where (10) and (11) are satisfied,

$$\left[\theta, \varphi \middle| D^{(t)}, P^{(t)} \right] = \left[\theta, \varphi \middle| D^{(t)} \right] \propto \left[D^{(t)} \middle| \theta, \varphi \right] \left[\theta \middle| \varphi \right] \left[\varphi \right].$$

Whilst the above is formulated under a Bayesian framework, I note that the data, $D^{(t)}$ and $P^{(t)}$, are ancillary for Y_t or $(\theta, \varphi)'$ under the assumptions laid out above. Under the conditionality principle, frequentist inference for Y_t or $(\theta, \varphi)'$ should be based on holding the data, $D^{(t)}$ and $P^{(t)}$, fixed (see, for example, Cox and Hinkley, 1974, page 31).

6. DATA OWNERSHIP

I also agree that data ownership and access is a key issue for NSOs and one where there is a generally lack of legislation and a supporting framework (page 30). The challenge is to unlock public good from privately collected data whilst protecting the commercial interests of the data custodians.

In many cases, commercial value is placed on primary and derived non-government data sets by their owners, since either the provision of such data is the basis of their business, or its possession is a significant element of competitive advantage. This raises the issue of how the NSO might acquire commercially valuable or sensitive data for statistical production, particularly if the statistics compete directly with information products created by the data owner or they compromise its market position. This issue is made more complex by the fact that there may be several parties with some form of commercial right in relation to a data set, either through ownership, possession or licensing arrangements.

Much Web content is also unstructured and ungoverned – the metadata describing its usage and provenance (origin, derivation, history, custody, and context) are either incomplete or incongruous. Indeed, the long-term reliability of Big Data sources may be an issue for ongoing statistical production. Reputable statistics for policy making and service evaluation are generally required for extended periods of time, often many years. However, large data sets from dynamic networks are volatile (and arguable static sources as well) – the data sources may change in character or disappear over time. This transience of data streams and sources does not sit comfortably with the reliability of statistical production and publication of meaningful time series.

With more statistics potentially available from the Web subject to different levels of biases and measurement errors at different points in time, what guidance can statisticians provide to report, connect and compare these statisticians over time and between different sources? As a minimum, the statistical profession should encourage the dissemination of these statistics to be accompanied by relevant meta data, for example, in the form of quality declarations and in accordance with Quality Frameworks (ABS, 2010; Brackstone, 1999; OECD, 2011) widely adopted by official statisticians.

7. A POSSIBLE ANALYSIS OF SATELLITE DATA TO PREDICT CROP YIELDS

To illustrate the potential analysis being developed in the ABS, I shall assume that equations (5), (6) and (7) are fulfilled by satellite data. Equation (6) is satisfied by choosing a random sample of observation units and collecting (ground truth) data on crop yields – the data are then integrated with satellite data to provide the 'training dataset'. Equation (7) is fulfilled as the coverage of satellite data is the same as the coverage for land parcels. Equation (5) may not hold for certain areas in Australia due to persistent cloud cover, as a result of moisture in the atmosphere, which may affect the type of crops being grown, or yields. This issue may, however, be by-passed by using traditional data collections e.g. statistical surveys, instead of using satellite data, for these areas.

Let the $\mathrm{N} \times 1$ vectors $\mathbf{M}_{\mathrm{t}}$, $\mathbf{m}_{\mathrm{t}}$ and $\mathbf{Q}_{\mathrm{t}}$ be the column vector of the crop yield, crop type and quantity harvestable respectively, for every observation unit of Australia.

Then, $\mathbf{M}_{\mathrm{t}} = \mathbf{Q}_{\mathrm{t}} * \mathbf{m}_{\mathrm{t}} = \mathbf{Exp}\left(\mathrm{Y}_{\mathrm{t}}\right) * \mathbf{m}_{\mathrm{t}}$, where $*$ denotes the Hadamard product, the $\mathrm{N} \times 1$ vector $\mathbf{Exp}\left(\mathrm{Y}_{\mathrm{t}}\right)$ has $exp\left(\mathrm{Y}_{\mathrm{it}}\right)$ as its i-th element, $\mathrm{Y}_{\mathrm{it}} = \log \mathrm{Q}_{\mathrm{it}}$ and Q_{it} is the i-th element of $\mathbf{Q}_{\mathrm{t}}$. Under the MAR assumptions made above, we can ignore $\mathbf{P}^{(\mathrm{t})}$ for predictive inference. That is,

$$\begin{aligned}\left[\mathrm{Y}_{\mathrm{t}}, \mathbf{m}_{\mathrm{t}} \middle| \mathrm{D}^{(\mathrm{t})}, \mathbf{P}^{(\mathrm{t})}\right] &= \left[\mathrm{Y}_{\mathrm{t}}, \mathbf{m}_{\mathrm{t}} \middle| \mathrm{D}^{(\mathrm{t})}\right] \\ &= \left[\mathrm{Y}_{\mathrm{t}} \middle| \mathbf{m}_{\mathrm{t}}, \mathrm{D}^{(\mathrm{t})}\right]\left[\mathbf{m}_{\mathrm{t}} \middle| \mathrm{D}^{(\mathrm{t})}\right].\end{aligned}$$

By assuming $\mathbf{m}_{\mathrm{t}}$ and $\mathrm{Y}_{\mathrm{t}} \mid \mathbf{m}_{\mathrm{t}}$ can be modelled by Dynamic Logistic Regression and Dynamic Linear models respectively, Tam and Clarke (2015b) provided results for the predictive distributions, $\left[\mathrm{Y}_{\mathrm{t}} \mid \mathbf{m}_{\mathrm{t}}, \mathrm{D}^{(\mathrm{t})}\right]$ and $\left[\mathbf{m}_{\mathrm{t}} \mid \mathrm{D}^{(\mathrm{t})}\right]$.

To illustrate the idea for predicting quantity, under the assumptions of this Section, (3) becomes

$$\begin{bmatrix} \mathrm{Y}_{\mathrm{ot}} \\ \mathrm{Y}_{\mathrm{rt}} \end{bmatrix} = \begin{bmatrix} \mathrm{Z}_{\mathrm{ot}} \\ \mathrm{Z}_{\mathrm{rt}} \end{bmatrix} \beta_{\mathrm{t}} + \begin{bmatrix} \mathrm{e}_{\mathrm{ot}} \\ \mathrm{e}_{\mathrm{rt}} \end{bmatrix} \qquad (12)$$

in which we have dropped the subscript 'B' to simplify notation. See Section 7.2 below for the choice of covariates and suggestions for improving the model in (12).

Assuming that

$$\begin{aligned}
&\mathbf{Y}_{\mathrm{t}} \mid \mathbf{Z}_{\mathrm{t}}, \boldsymbol{\beta}_{\mathrm{t}} \sim N\left(\mathbf{Z}_{\mathrm{t}}\boldsymbol{\beta}_{\mathrm{t}}, \boldsymbol{\Sigma}_{\mathrm{t}}\right)\\
&\boldsymbol{\beta}_{\mathrm{t}} = \mathbf{H}\boldsymbol{\beta}_{\mathrm{t}-1} + \boldsymbol{\varepsilon}_{\mathrm{t}} \;\;,\;\; \boldsymbol{\beta}_{\mathrm{t}} \perp \mathbf{Z}_{\mathrm{t}}\\
&\boldsymbol{\beta}_1 \sim N\left(\boldsymbol{\beta}_0, \boldsymbol{\Omega}_{\boldsymbol{\beta}_0}\right)\\
&\boldsymbol{\varepsilon}_{\mathrm{t}} \sim \text{independent } N\left(\mathbf{0}, \boldsymbol{\Omega}_{\mathrm{t}}\right) \;\;,\;\; \boldsymbol{\varepsilon}_{\mathrm{t}} \perp \mathbf{D}^{(\mathrm{t})}
\end{aligned} \tag{13}$$

and $\boldsymbol{\Omega}_{\mathrm{t}}$ and $\boldsymbol{\Sigma}_{\mathrm{t}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathrm{oot}} & 0 \\ 0 & \boldsymbol{\Sigma}_{\mathrm{rrt}} \end{pmatrix}$ are known, the predictive distribution of the total yield of a particular crop (Tam and Clarke, 2015b) is $\mathbf{1}_{\mathrm{o}}'\mathbf{Q}_{\mathrm{ot}} + \mathbf{1}_{\mathrm{r}}'\mathbf{Exp}\left(\hat{\mathbf{Y}}_{\mathrm{rt}}\right)$, where

$$\begin{aligned}
\hat{\mathbf{Y}}_{\mathrm{rt}} &\sim N\left(\mathbf{Z}_{\mathrm{rt}}\hat{\boldsymbol{\beta}}_{\mathrm{t|t}}\;, \boldsymbol{\Sigma}_{\mathrm{rrt}} + \mathbf{Z}_{\mathrm{rt}}\boldsymbol{\Omega}_{\mathrm{t|t}}\mathbf{Z}_{\mathrm{rt}}'\right)\\
\hat{\boldsymbol{\beta}}_{\mathrm{t|t}} &= \mathbf{H}\hat{\boldsymbol{\beta}}_{\mathrm{t-1|t-1}} + \boldsymbol{\Omega}_{\mathrm{t|t}}\mathbf{Z}_{\mathrm{ot}}'\boldsymbol{\Sigma}_{\mathrm{oot}}^{-1}\left(\mathbf{Y}_{\mathrm{ot}} - \mathbf{Z}_{\mathrm{ot}}\hat{\boldsymbol{\beta}}_{\mathrm{t-1|t-1}}\right)\\
\boldsymbol{\Omega}_{\mathrm{t|t}} &= \left(\boldsymbol{\Omega}_{\mathrm{t|t-1}}^{-1} + \mathbf{Z}_{\mathrm{ot}}'\boldsymbol{\Sigma}_{\mathrm{oot}}^{-1}\mathbf{Z}_{\mathrm{ot}}\right)^{-1}\\
\boldsymbol{\Omega}_{\mathrm{t|t-1}} &= \mathbf{H}\boldsymbol{\Omega}_{\mathrm{t-1|t-1}}\mathbf{H}' + \boldsymbol{\Omega}_{\mathrm{t}} \;.
\end{aligned} \tag{14}$$

Here $\mathbf{Exp}\left(\hat{\mathbf{Y}}_{\mathrm{rt}}\right)$ denotes the vector with $exp\left(\hat{\mathrm{Y}}_{\mathrm{irt}}\right)$ as its i-th element, $\hat{\mathrm{Y}}_{\mathrm{irt}}$ is the i-th element of $\hat{\mathbf{Y}}_{\mathrm{rt}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{t|t}}$ denotes the posterior mean of $\boldsymbol{\beta}_{\mathrm{t}}$ given $\mathbf{D}^{(\mathrm{t})}$, and $\boldsymbol{\Omega}_{\mathrm{t|t}}$ is the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{t|t}}$.

Note the above methodology may be adapted to a ‘design-assisted’ approach (Särndal *et al.*, 1992) for estimating finite population parameters using the following heuristic argument. From (3), the Generalised Regression Estimator for the total yield, $\mathbf{1}'\mathbf{Y}_{\mathrm{t}}$, is

$$\mathrm{e}_{\mathrm{ot}}\left(\mathbf{Y}_{\mathrm{t}}\right) + \left\{\mathbf{1}'\mathbf{Z}_{\mathrm{t}} - \mathrm{e}_{\mathrm{ot}}\left(\mathbf{Z}_{\mathrm{t}}\right)\right\}\widehat{\boldsymbol{\beta}}_{\mathrm{Dt}}$$

where $\mathrm{e}_{\mathrm{ot}}\left(\mathbf{Y}_{\mathrm{t}}\right)$, $\mathrm{e}_{\mathrm{ot}}\left(\mathbf{Z}_{\mathrm{t}}\right)$ are the Horvitz-Thompson estimators of $\mathbf{Y}_{\mathrm{t}}$ and $\mathbf{Z}_{\mathrm{t}}$ respectively, and $\widehat{\boldsymbol{\beta}}_{\mathrm{Dt}}$ is the design based estimator of $\boldsymbol{\beta}_{\mathrm{t}}$ at time t (Särndal *et al.*, 1992). Following Wright (1983), even though $\hat{\boldsymbol{\beta}}_{\mathrm{t|t}}$ is not asymptotically design unbiased, we may use it for $\widehat{\boldsymbol{\beta}}_{\mathrm{Dt}}$.

Likewise, denoting $\sigma\left(Z'_{it}\gamma_t\right)=\left[1+exp\left(-Z'_{it}\gamma_t\right)\right]^{-1}$ as the logistic sigmoid for observation i at time t, and assuming m_{it} ~ independent Binomial Logistic $\left(\sigma\left(Z'_{it}\gamma_t\right)\right)$, or

$$\begin{aligned}
\begin{bmatrix} m_{ot} \\ m_{rt} \end{bmatrix} &= \begin{bmatrix} \sigma\left(Z_{ot}\gamma_t\right) \\ \sigma\left(Z_{rt}\gamma_t\right) \end{bmatrix} \\
\gamma_t &= H\gamma_{t-1}+\varepsilon_t \ \ , \ \ \gamma_t \perp Z_t \\
\gamma_1 &\sim \mathcal{N}\left(\gamma_0, \Xi_{\gamma_0}\right) \\
\varepsilon_t &\sim \text{independent } N\left(0, \Xi_t\right) \ \ , \ \ \varepsilon_t \perp D^{(t)}
\end{aligned} \tag{15}$$

where $m_{ot}=\left(m_{1t},\ldots,m_{ot}\right)'$ and $\sigma\left(Z_{ot}\gamma_t\right)=\left(\sigma\left(Z'_{1t}\gamma_t\right),\ldots,\sigma\left(Z'_{nt}\gamma_t\right)\right)'$ etc. and Ξ_t is known, then (Tam and Clarke, 2015b)

$$m_{it} \mid D^{(t)} \approx \text{independent Binomial Logistic}\left(\sigma\left(Z'_{it}\hat{\gamma}_{t|t}\right)\right) \tag{16}$$

for unobserved units $i=1,\ldots,r_t$,

where $$\hat{\gamma}_{t|t}=H\,\hat{\gamma}_{t-1|t-1}+\Sigma^{-1}_{t|t-1}\left\{Z'_{ot}m_{ot}-Z'_{ot}\sigma\left(Z'_{ot}\hat{\gamma}_{t|t}\right)\right\}$$

and $$\Sigma_{t|t-1}=\Sigma_{t-1|t-1}+\Xi_t \ .$$

7.1 Statistical computing issues

The examples shown above make the unrealistic assumptions that quantities like Σ_t, Ω_t and Ξ_t are known. In reality they are not and have to be estimated by the observed data. To make the estimation task more manageable, one can consider modelling the unknown quantities as follows

$$\begin{aligned}
\Sigma_t &= \lambda_t(\Sigma)\,\Sigma \\
\Omega_t &= \lambda_t(\Omega)\,\Omega \\
\Xi_t &= \lambda_t(\Xi)\,\Xi
\end{aligned}$$

where the scalars $\lambda_t(\Sigma)$, $\lambda_t(\Omega)$, $\lambda_t(\Xi)>0$ follow an uninformative prior,

$$\Sigma \sim W^{-1}\left(\Sigma_0, \nu_\Sigma\right) \ \ , \ \ \Omega \sim W^{-1}\left(\Omega_0, \nu_\Omega\right) \ \ , \ \ \Xi \sim W^{-1}\left(\Xi_0, \nu_\Xi\right)$$

and W^{-1} denotes the Inverse-Wishart distribution.

Let $\Theta_t = \{\theta, \phi, \varphi, \lambda_t(\Sigma), \lambda_t(\Omega), \lambda_t(\Xi), \Sigma, \Omega, \Xi\}$ and may also include H if it is not known. Assuming (4) and (9) are fulfilled, then the posterior distribution of Θ_t ,

$$\left[\Theta_t \mid D^{(t)}\right] \propto \left[D^{(t)} \mid \Theta_t\right]\left[\Theta_t\right],$$

i.e. likelihood times the prior. 'Maximum a posteriori' estimates of Θ_t can be derived using the EM algorithm – see Haykin (2001, Chapter 5) and also Strickland *et al.* (2009, 2011) for efficient estimation applied to satellite data.

Alternatively, the predictive distribution, $\left[M_t \mid D^{(t)}\right]$, where $M_t = E(Y_t) * m_t$ as before, can be derived using Monte Carlo via the method of composition as follows. From

$$\begin{aligned}\left[Y_t, m_t, \Theta_t \middle| D^{(t)}\right] &= \left[Y_t, m_t \middle| D^{(t)}, \Theta_t\right]\left[\Theta_t \middle| D^{(t)}\right] \\ &= \left[Y_t \middle| m_t, D^{(t)}, \Theta_t\right]\left[m_t \middle| D^{(t)}, \Theta_t\right]\left[\Theta_t \middle| D^{(t)}\right],\end{aligned}$$

one can use the *LibBi* software as outlined in Murray (2015) to draw J samples $\Theta_t^1, \ldots, \Theta_t^J$ from $\left[D^{(t)} \mid \Theta_t\right]\left[\Theta_t\right]$. Using these values and equations (14) and (16), we obtain samples $Y_t^1, \ldots, Y_t^J$ from $N\left(Z_{rt}\hat{\beta}_{t|t}, \Sigma_{rrt} + Z_{rt}\Omega_{t|t}Z'_{rt}\right)$ and $m_t^1, \ldots, m_t^J$ from $\left[m_t \middle| D^{(t)}, \Theta_t\right]$ respectively, where the i-th element of the vector m_t follows a Binomial Logistic Regression model with logistic sigmoid $\sigma\left(Z'_{ti}\hat{\gamma}_t\right)$, from which a sample of $M_t^1, \ldots, M_t^J$ can be obtained for Monte Carlo inference on $M_t \mid D^{(t)}$. Strickland *et al.* (2013) has also developed a Python package, pyMCMC, for fast multivariate state space modelling, which is scheduled for release in June, 2015.

7.2 Choosing covariates and improving the model fit

There is a huge literature in predicting crop yields, see for example, Johnson (2014) and the references therein. A review of the methodology is provided in Lobell (2013). Based on the science of crops, most of these use the Normalised Difference Vegetation Index (NDVI), or Enhanced Vegetation Index (EVI), which are simple functions of the near-infrared radiation and visible radiation, and other variables like soil moisture, land surface temperature etc. available from satellites and other sources are included as covariates. Stress Index as a covariate derived from thermal time and crop phenology both from remote sensing (Idso *et al.*, 1981; Jackson *et al.*, 1983; Rodriguez *et al.*, 2005) as well as directly modelled from a biophysical crop model (Potgieter *et al.*, 2005; Potgieter and Hammer, 2006) has been proposed. In addition, evapotranspiration derived from EVI and Global Vegetation Moisture Index has been suggested as covariates (Guerschman *et al.*, 2009). These covariates can be incorporated in an obvious way into the State Space Model described above, although care has to be exercised to ensure there is no collinearity issue, or model over-fitting.

Becker-Reshef *et al.* (2010) fitted a simple regression model using county yield statistics as response variables and NDVI as explanatory variables, and use it to predict yields. Newlands *et al.* (2014) extends this work by employing a multivariate regression model using NDVI and agro-climate data as covariates. In addition, their model also allows a lag-1 autoregressive term for crop yields and the coefficients to vary over time and space, although no stochastic relationships between these coefficients were exploited. Priors on the parameters of the multivariate regression model were constructed using residual bootstrapping (Bornn and Zidek, 2012). The State Space Modelling advocated in this paper can be regarded as an extension of the methodology developed by Newlands *et al.* (2014).

Where the model defined (13) does not adequately predict crop quantities, the following model may be considered:

$$\mathbf{Y}_t \mid \mathbf{Z}_t, \boldsymbol{\beta}_t, \mathbf{F}, \boldsymbol{\alpha}_t \;\sim\; N\left(\mathbf{F}\boldsymbol{\alpha}_t + \mathbf{Z}_t\boldsymbol{\beta}_t \,, \boldsymbol{\Sigma}_t\right)$$

$$\begin{bmatrix} \boldsymbol{\alpha}_t \\ \boldsymbol{\beta}_t \end{bmatrix} = \begin{bmatrix} \mathbf{H}_1\boldsymbol{\alpha}_{t-1} \\ \mathbf{H}_2\boldsymbol{\beta}_{t-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_{1t} \\ \boldsymbol{\varepsilon}_{2t} \end{bmatrix} \;,\; \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t \perp \mathbf{Z}_t$$

$$\boldsymbol{\varepsilon}_t = \begin{bmatrix} \boldsymbol{\varepsilon}_{1t} \\ \boldsymbol{\varepsilon}_{2t} \end{bmatrix} \sim \text{independent}\, N\left(0, \boldsymbol{\Omega}_t\right) \;,\; \boldsymbol{\varepsilon}_t \perp \mathbf{D}^{(t)}$$

$$\boldsymbol{\Omega}_t = \begin{pmatrix} \boldsymbol{\Omega}_{1t} & 0 \\ 0 & \boldsymbol{\Omega}_{2t} \end{pmatrix}.$$

In other words, the time-variant fixed effects, $\mathbf{F}\boldsymbol{\alpha}_t$, is used to ‘sweep’ up any missing covariates in the modelling. This approach is akin to using random slopes in multi-level modelling (Snijders and Bosker, 1999 – Chapter 5) and is also known as Generalised Linear Mixed Model. A similar approach may be adopted for the model described in equation (14). The suggested approach, however, would require large sample sizes, as well as longer time series for accurate and precise estimation.

7.3 Concluding remarks

In developing the above models and building the training data set for analyses, I found that I have to involve crop scientists (or more generally "domain experts" – page 26 of the AAPOR Report), statisticians and computer scientists, supporting the comment that a multi-disciplinary team is required to harness opportunities, and addressing challenges, from Big Data. New skill sets are required to integrate ground truth data with satellite data.

Recommendation 1 of the AAPOR Report (page 2) says:

> "Survey and Big Data are complementary data sources and not competing data sources. There are differences between the approaches, but this should be seen as an advantage rather than a disadvantage".

This paper outlines an approach to combine the strength of Big Data with survey data – which has been regarded as the 'gold standard' for collecting data to make valid statistical inference – for predicting crop yields. The basic ideas are to use the Big Data and other auxiliary sources to calibrate the response variables, and to apply State Space Modelling to solve finite population inference problems. However, this is possible because the population covers by satellite imagery is identical to the population of land parcels, and the missing covariates problem is by-passed by relying on the traditional survey methods of estimation in those areas without satellite data e.g. missing data due to clouds. The efficacy of the approach will be tested using the training data set that is being built in the ABS. I hope to be able to report the outcome of the analyses, successful or otherwise, in the future elsewhere.

Once again, I congratulate the AAPOR Task Force for providing an excellent Report.

ACKNOWLEDGEMENTS

I thank Professor Natalie Shlomo for inviting me to comment on the AAPOR Report. I would also like to thank Dr Andries Potgieter and Dr Juan Guerschman for their insights on crop science, Professor Kerrie Mengersen for her many helpful suggestions to improve the paper, and my colleagues Jennifer Marley, Daniel Elazar and Sybille McKeown for useful discussions.

REFERENCES

Australian Bureau of Statistics (2010a) *The ABS Data Quality Framework.* <https://www.nss.gov.au/dataquality/aboutqualityframework.jsp>

Becker-Reshef, I.; Vermote, E.; Lindeman, M. and Justice, C. (2010) "A Generalised Regression-Based Model for Forecasting Winter Wheat Yields in Kansas and Ukraine Using MODIS Data", *Remote Sensing of Environment*, 114(6), pp. 1312–1323.

Bornn, L. and Zidek, J.V. (2012) "Efficient Stabilization of Crop Yield Prediction in the Canadian Prairies", *Agricultural and Forest Meteorology*, 152, pp. 223–232.

Brackstone, G. (1999) "Managing Data Quality in a Statistical Agency", *Survey Methodology*, 25(2), pp. 139–149.

Chambers, R.L. and Clark, R.G. (2012) *An Introduction to Model-Based Survey Sampling with Applications*, Oxford University Press, London.

Couper, M.P. (2013) "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys", *Survey Research Methods*, 7(3), pp. 145–156. <https://ojs.ub.uni-konstanz.de/srm/article/view/5751/5289>

Cox, P.R. and Hinkley, D.V. (1974) *Theoretical Statistics*, Chapman and Hall, London.

Denham, R.J.; Falk, M.G. and Mengersen, K.L. (2011) "The Bayesian Conditional Independence Model for Measurement Error: Applications in Ecology", *Environmental and Ecological Statistics*, 18(2), pp. 239–255.

Fan, J.; Han, F. and Liu, H. (2014) "Challenges of Big Data Analysis", *National Science Review*, 1(2), pp. 293–314.

Guerschman, J.P.; Van Dijk, A.I.; Mattersdorf, G.; Beringer, J.; Hutley, L.B.; Leuning, R.; Pipunic, R.C. and Sherman, B.S. (2009) "Scaling of Potential Evapotranspiration with MODIS Data Reproduces Flux Observations and Catchment Water Balance Observations Across Australia", *Journal of Hydrology*, 369(1–2), pp. 107–119.

Harford, T. (2014) "Big Data: A Big Mistake?", *Significance*, 11(5), pp. 14–19.

Haykin, S.S. (ed.) (2001) *Kalman Filtering and Neural Networks*, John Wiley & Sons, Inc., New York.

Iacus, S.M. (2014) "Big Data or Big Fall? The Good, the Bad and the Ugly and the Missing Role of Statistics", *Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation*, 5(1), pp. 4–11.

Idso, S.B.; Reginato, R.J.; Jackson, R.D. and Pinter, P.J. (1981) "Measuring Yield-Reducing Plant Water Potential Depressions in Wheat by Infrared Thermometry", *Irrigation Science*, 2(4), pp. 205–212.

Jackson, R.D.; Slater, P.N. and Pinter, P.J. (1983) "Discrimination of Growth and Water Stress in Wheat by Various Vegetation Indices Through Clear and Turbid Atmospheres", *Remote Sensing of Environment*, 13(3), pp. 187–208.

Japec, L.; Kreuter, F.; Berg, M.; Biemer, P.; Decker, P.; Lampe, C.; Lane, J.; O'Neil, C. and Usher, A. (2015) *American Association for Public Opinion Research (AAPOR) Report on Big Data.* <http://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15.pdf>

Johnson, D.M. (2014) "An Assessment of Pre- and Within-Season Remotely Sensed Variables for Forecasting Corn and Soybean Yields in the United States", *Remote Sensing of Environment*, 141, pp. 116–228.

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data*, Second Edition, Wiley, New York.

Lobell, D.B. (2013) "The Use of Satellite Data for Crop Yield Gap Analysis", *Field Crops Research*, 143, pp. 56–64.

Murray, L.M. (2015) *Bayesian State-Space Modelling on High-Performance Hardware using LibBi.* <http://arxiv.org/pdf/1306.3277v1.pdf>

Newlands, N.K.; Zamar, D.S.; Kouadio, L.A.; Zhang, Y.; Chipanshi, A.; Potgieter, A.; Toure, S. and Hill, H.S.J. (2014) "An Integrated, Probabilistic Model for Improved Seasonal Forecasting of Agricultural Crop Yield Under Environmental Uncertainty", *Frontiers in Environmental Science*, 2, pp. 1–21. <http://journal.frontiersin.org/article/10.3389/fenvs.2014.00017/full>

OECD (2011) *Quality Dimensions, Core Values for OECD Statistics and Procedures for Planning and Evaluating Statistical Activities*. <http://www.oecd.org/std/21687665.pdf>

Potgieter, A.B.; Hammer, G.L.; Doherty, A. and de Voil, P. (2005) "A Simple Regional-Scale Model for Forecasting Sorghum Yield Across North-Eastern Australia", *Agriculture and Forest Meteorology*, 132(1–2), pp. 143–153.

Potgieter, A.B. and Hammer, G.L. (2006) *Oz-Wheat: A Regional-Scale Crop Yield Simulation Model for Australian Wheat*, Information Series, Queensland Department of Primary Industries and Fisheries, Brisbane.

Puza, B. (2013) *Lectures on Bayesian Statistics*, Unpublished manuscript, Research School of Finance, Actuarial Studies and Applied Statistics, Australian National University.

Rodriguez, D.; Fitzgerald, G.J.; Belford, R. and Christensen, L. (2006) "Detection of Nitrogen Deficiency in Wheat From Spectral Reflectance Indices and Basic Crop Eco-Biophysiological Concepts", *Australian Journal of Agricultural Research*, 57(7), pp. 781–789.

Rubin, D.B. (1976) "Inference and Missing Data", *Biometrika*, 63(3), pp. 581–592.

Särndal, C.-E.; Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Snijders, T. and Bosker, R. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*, Sage, London.

Snijders, T.A.B. and Bosker, R.J. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, SAGE Publications Ltd, London.

Speed, T. (2014) *Data Science, Big Data and Statistics: Can We All Live Together?* Chalmers Initiative Seminar on Big Data. <http://www.chalmers.se/en/areas-of-advance/ict/events/Documents/Terry%20Speed_Data%20Science,%20Big%20Data%20and%20Statistics%20-%20Can%20We%20All%20Live%20Together.pdf>

Strickland, C.M.; Turner, I.W.; Denham, R.J. and Mengersen, K.L. (2009) "Efficient Bayesian Estimation of Multivariate State Space Models", *Computational Statistics and Data Analysis*, 53(12), pp. 4116–4125.

Strickland, C.M.; Simpson, D.P.; Turner, I.W.; Denham, R.J. and Mengersen, K.L. (2011) "Fast Bayesian Analysis of Spatial Dynamic Factor Models for Multi-Temporal Remotely Sensed Imagery", *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60, pp. 109–124.

Strickland, C.M.; Denham, R.J.; Alston, C.L. and Mengersen, K.L. (2013) "PyMCMC : A Python Package for Bayesian Estimation using Markov Chain Monte Carlo", in C.L. Alston, K.L. Mengersen and A.N. Pettitt (eds.), *Case Studies in Bayesian Statistical Modelling and Analysis,* John Wiley, London, pp. 421–460.

Tam, S.M. (1987) "Analysis of Repeated Surveys Using a Dynamic Linear Model", *International Statistical Review*, 55, pp. 67–73.

Tam, S.M. and Clarke, F. (2015a) "Big Data, Official Statistics and Some Initiatives of the Australian Bureau of Statistics", *International Statistical Review* (to appear).

Tam, S.M. and Clarke, F. (2015b) "Big Data, Statistical Inference and Official Statistics", *Methodology Research Papers*, cat. no. 1351.0.55.054, Australian Bureau of Statistics, Canberra. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.054>

Wright, R.L. (1983) "Finite Population Sampling with Multivariate Auxiliary Information", *Journal of the American Statistical Association*, 78(384), pp. 879–884.

Wu, L. (2010) *Mixed Effect Models for Complex Data*, CRC Press, Florida.

All URLs last viewed on Monday 15 June 2015

FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** The ABS website is the best place for data from our publications and information about the ABS.

LIBRARY A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our website for a list of libraries.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE	1300 135 070
EMAIL	client.services@abs.gov.au
FAX	1300 135 211
POST	Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS www.abs.gov.au

1351.0.55.056 • RESEARCH PAPER: A STATISTICAL FRAMEWORK FOR ANALYSING BIG DATA

Produced by the Australian Bureau of Statistics