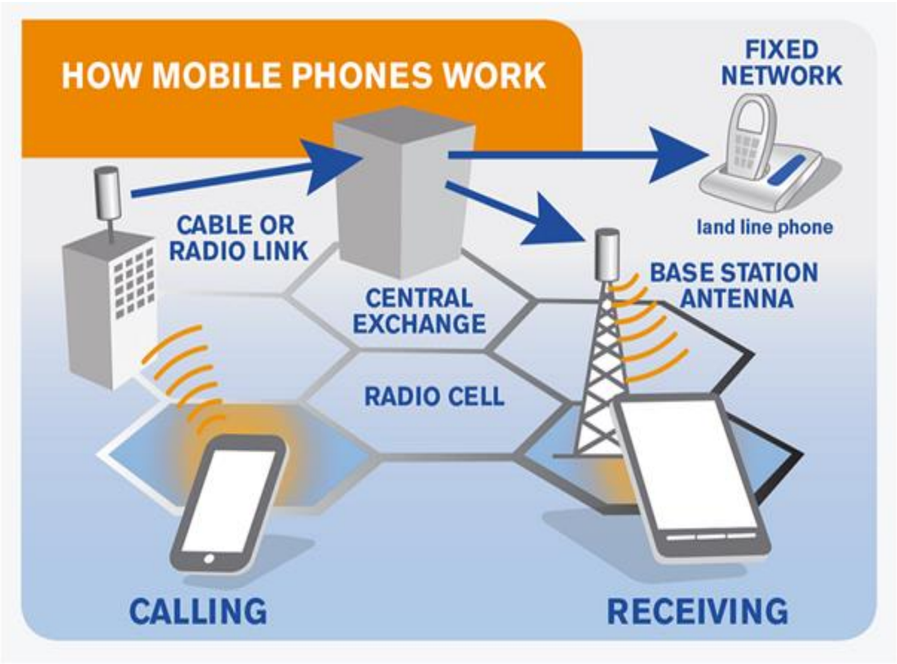
A Statistical Framework for Analysing Big Data

Global Conference on Big Data for Official Statistics

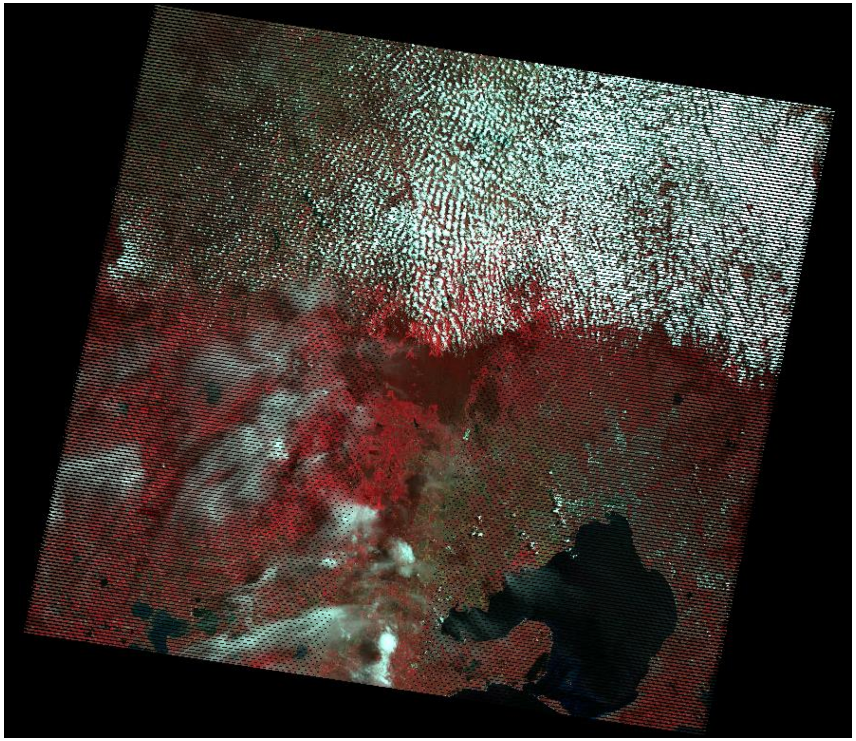
20-22 October, 2015

by S Tam, Chief Methodologist

Australian Bureau of Statistics



Reference – ITU EMF Guide 2014



Big Data (BD) Issues

- Characterise BD with big “N”, big “p” and big “t”
- Sampling error
  - Reduced by increasing the sample size
- Bias
  - Coverage bias - Big Data population is not the population of interest
  - Self selection bias – are their views representative of the “silent” population segments?
  - Representation bias – multiple representation
  - Measurement error – Are the data related to the concept of interest?
  - Increasing the sample size does NOT reduce non-sampling errors

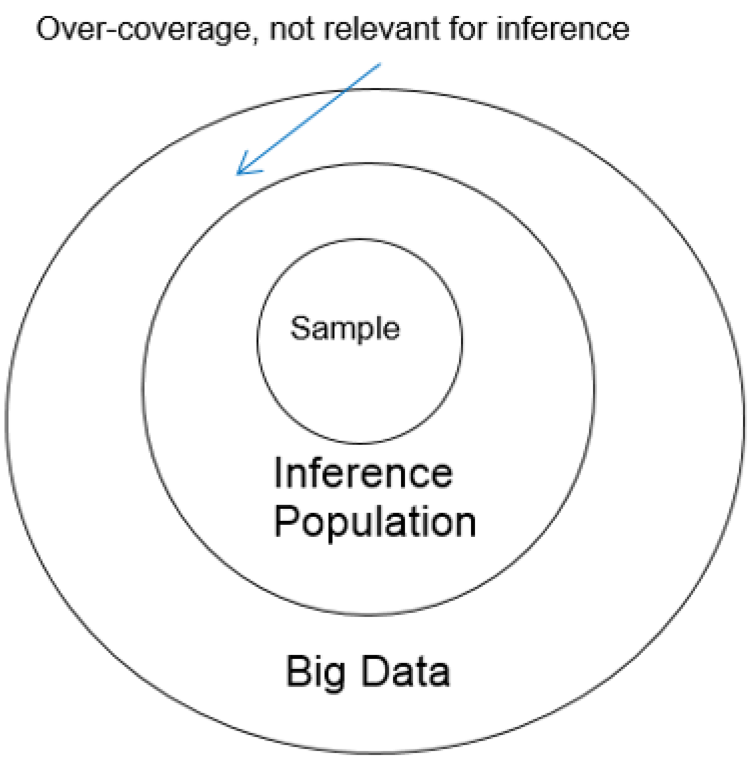
Methods overview

- Domain (e.g. crop) modelling
- Machine learning methods
  - Decision Trees
  - Artificial Neural Networks
  - Support Vector Machines
  - Nearest Neighbour
  - Ensemble classifiers
- Statistical modelling
  - Spatial-Temporal models
    - Use “space” and “time” information
    - Use in ABS modelling for satellite imagery and simulated phone data

ABS method overview

Two-Step Approach – Calibration and Prediction

- Use a sample to calibrate the Big Data (treated as “covariates”) using ground truths/measurements
- Calibrate using a linear model with time varying coefficients – Dynamic Model
- Estimate parameters (using Frequentist/Bayesian approaches)
- Predict the non-sampled values using the covariates



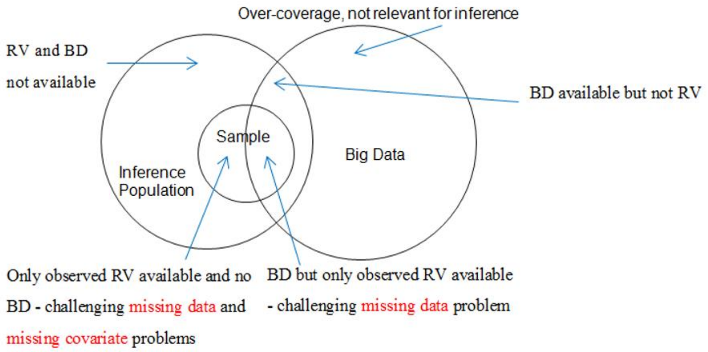
$$\begin{bmatrix} \mathbf{Y}_{ot} \\ \mathbf{Y}_{rt} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_{ot} \\ \mathbf{Z}_{rt} \end{bmatrix} \boldsymbol{\beta}_t + \begin{bmatrix} \mathbf{e}_{ot} \\ \mathbf{e}_{rt} \end{bmatrix}$$

Selectivity - Missing Covariate challenges

The selectivity bias issue

- When the Big Data population only intersects with the inference population, Big Data covariates are missing
  - Two problems
    - Bias in estimating beta
    - Covariates are missing
  - Can we just ignore the selectivity bias issue?

Big Data bias and statistical modelling



Note: RV = Response variable from "ground truths"; BD = Big Data

$$\begin{bmatrix} \mathbf{Y}_{O_B t} \\ \mathbf{Y}_{O_{\bar{B}} t} \\ \mathbf{Y}_{rt} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_{O_B t} \\ \mathbf{Z}_{O_{\bar{B}} t} \\ \mathbf{Z}_{rt} \end{bmatrix} \boldsymbol{\beta}_t + \begin{bmatrix} \mathbf{e}_{O_B t} \\ \mathbf{e}_{O_{\bar{B}} t} \\ \mathbf{e}_{rt} \end{bmatrix}$$

When can selectivity be ignored?

Inference on finite population values

- Sampling and missing data processes can be ignored if “these processes are not dependent on the unobserved population values”
  - Fulfilled if the training data set is selected by probability sampling
  - There is no missing data or the missing data is “missing at random
- Big Data process can be ignored if “this process is not dependent on the missing covariates”
  - This condition is difficult to check
  - If condition not satisfied, modelling for the missing covariates – a difficult task – will be needed.

Modelling the missing processes

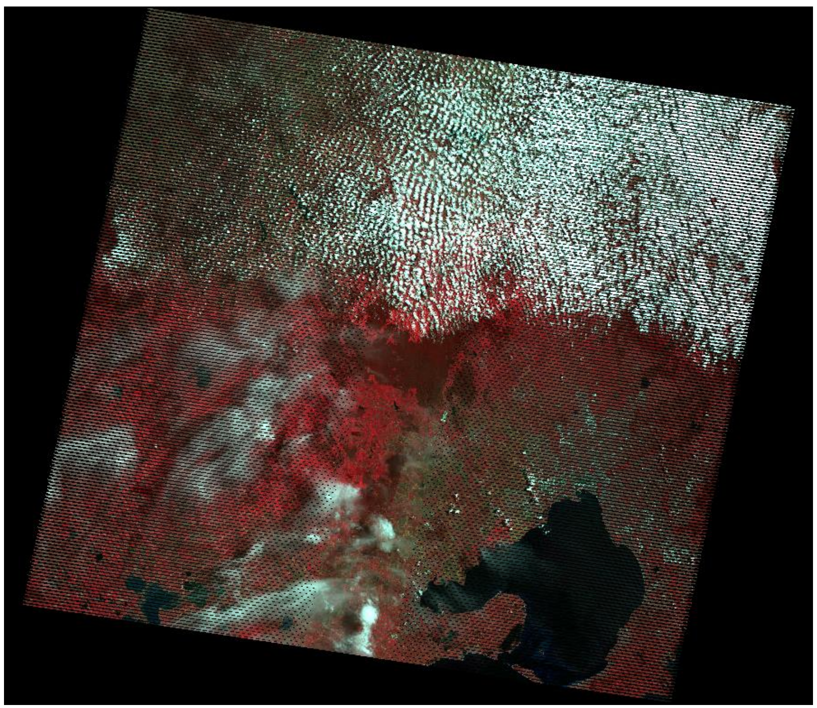
- 3 processes at play at time t, with reference to the inference population
  - Sampling process – $I_{ti} = \{0,1\}$
  - Missing observation process, $R_{ti} = \{0,1\}$
  - Big Data process, $\mathcal{R}_{ti} = \{0,1\}$
- Data on the censoring processing
  - $P^{(t)}_1$= Data on I_i and R_i for I = 1,.., t
  - $P^{(t)}_2$= Data on $\mathcal{R}_i$

$$\left[\mathbf{P}_1^{(t)} \,\middle|\, \mathbf{Y}_t, \mathbf{D}^{(t)}, \mathbf{P}_2^{(t)}, \Theta\right] = \left[\mathbf{P}_1^{(t)} \,\middle|\, \mathbf{D}^{(t)}, \mathbf{P}_2^{(t)}\right]$$

$$\left[\mathbf{P}_2^{(t)} \,\middle|\, \mathbf{Y}_t, \mathbf{D}^{(t)}, \mathbf{D}_c^{(t)}, \Theta\right] = \left[\mathbf{P}_2^{(t)} \,\middle|\, \mathbf{Y}_t, \mathbf{D}^{(t)}\right]$$

Predicting crop types from Satellite Imagery

Satellite imagery



Data at time = t

Big Data

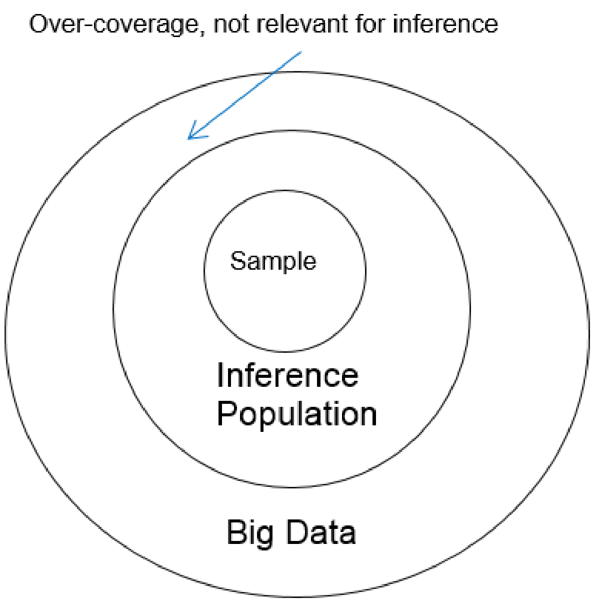
- reflectance data from 7 frequency bands from satellite imagery

Band1	Band2	Band3	Band4	Band5	Band6	Band7
514	745	888	1908	2112	2233	1356
584	708	953	1763	1940	2233	1378
532	727	985	1872	1961	2233	1290
550	764	985	1981	2197	2233	1489
550	764	969	1981	2069	2233	1356
550	745	985	1945	2048	2233	1312
550	690	921	1799	2197	2182	1512
584	727	888	1727	2175	2182	1489
584	708	888	1763	2154	2130	1512
532	727	904	1763	2133	2130	1489

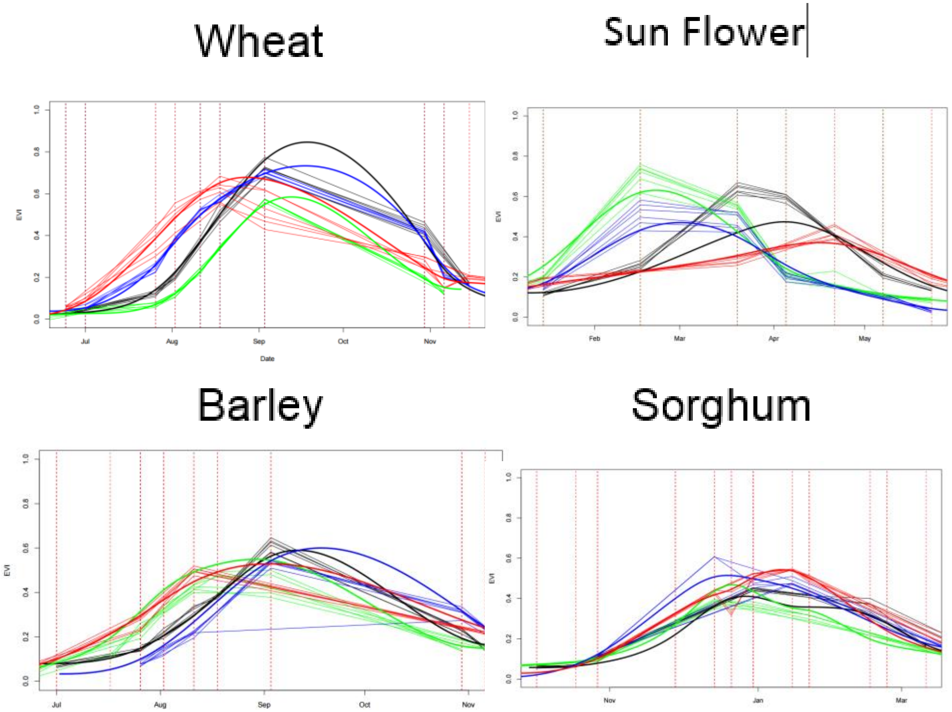
- Statistical models require ground truths/measurements

Some interesting observations

Satellite Imagery – no under-coverage (missing covariates) issues



Enhanced vegetation index – EVI plotted over the growing season

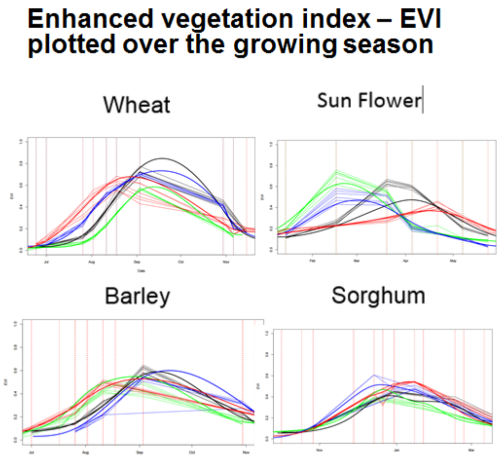


$$EVI = G \times \frac{(NIR - RED)}{(NIR + C1 \times RED - C2 \times Blue + L)}$$

L=1, C1 = 6, C2 = 7.5, and G (gain factor) = 2.5.

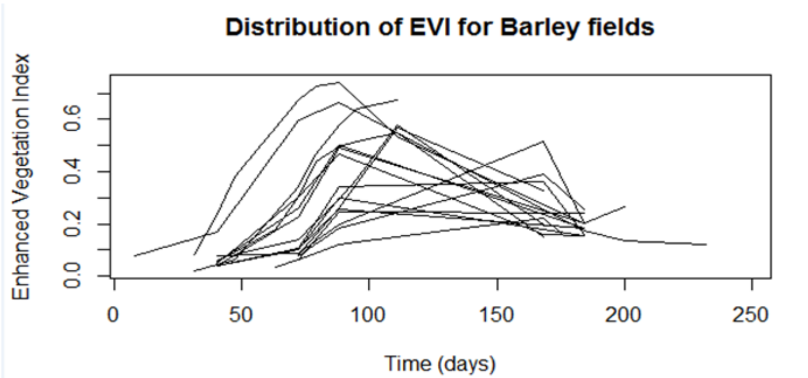
What covariates to use?

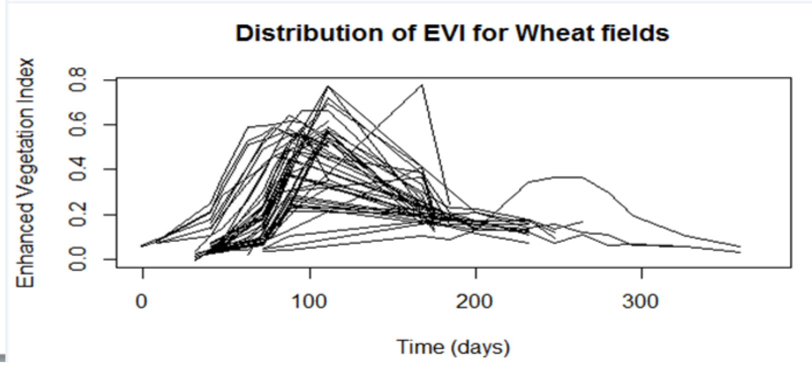
- This is where crop science comes in
  - Possible covariate curves
    - Land surface temperature curve
    - Moisture curve
    - Grow curves etc.
  - Want to use the whole curve
  - Need a method to pick finite points
    - Functional Data Analysis
      - Pick the points which explains most of the variation of the curve
        - » Functional Data Analysis



Data, Static Models and Accuracy

The training data



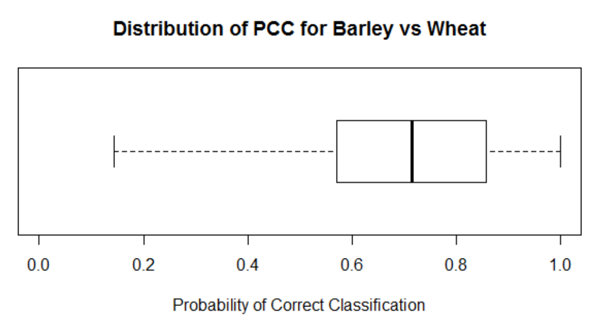


Models and accuracy

$$\begin{bmatrix} \mathbf{p}_{it} \\ \mathbf{p}_{jt} \end{bmatrix} \cong \begin{bmatrix} \left[1 + exp\left(-\mathbf{Z}'_{it}\gamma_t\right)\right]^{-1} \\ \left[1 + exp\left(-\mathbf{Z}'_{jt}\gamma_t\right)\right]^{-1} \end{bmatrix}$$

****

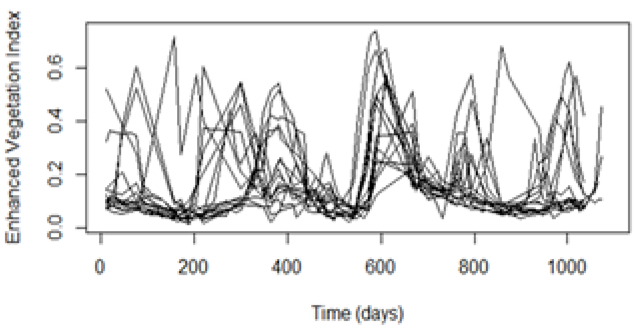
$$\begin{bmatrix} \mathbf{Y}_{ot} \\ \mathbf{Y}_{rt} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_{ot} \\ \mathbf{Z}_{rt} \end{bmatrix} \beta_t + \begin{bmatrix} \mathbf{e}_{ot} \\ \mathbf{e}_{rt} \end{bmatrix}$$



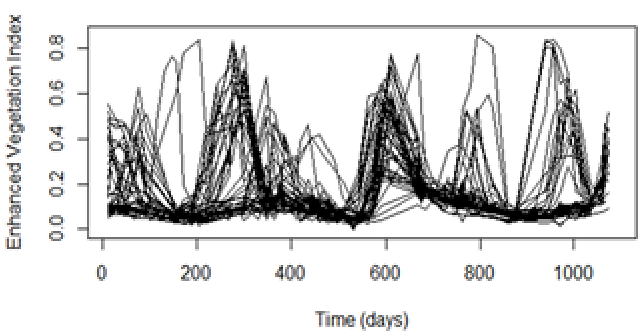
Temporal Modelling

Data over time

Distribution of EVI for Barley fields



Distribution of EVI for Wheat fields



Dynamic models

$$\begin{bmatrix} \mathbf{p}_{it} \\ \mathbf{p}_{jt} \end{bmatrix} \cong \begin{bmatrix} \left[1 + exp\left(-\mathbf{z}'_{it}\gamma_t\right)\right]^{-1} \\ \left[1 + exp\left(-\mathbf{z}'_{jt}\gamma_t\right)\right]^{-1} \end{bmatrix}$$

$$\gamma_t = \mathbf{H}\gamma_{t-1} + \varepsilon_t$$

$$* * * * * * *$$

$$\begin{bmatrix} \mathbf{Y}_{ot} \\ \mathbf{Y}_{rt} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_{ot} \\ \mathbf{Z}_{rt} \end{bmatrix} \beta_t + \begin{bmatrix} \mathbf{e}_{ot} \\ \mathbf{e}_{rt} \end{bmatrix}$$

$$\beta_t = \mathbf{M}\beta_{t-1} + \varepsilon_t$$

The Algorithms

Dynamin Linear Model

$$\begin{bmatrix} \mathbf{Y}_{ot} \\ \mathbf{Y}_{rt} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_{ot} \\ \mathbf{Z}_{rt} \end{bmatrix} \boldsymbol{\beta}_t + \begin{bmatrix} \mathbf{e}_{ot} \\ \mathbf{e}_{rt} \end{bmatrix}$$

$$\boldsymbol{\beta}_t = \mathbf{M}\boldsymbol{\beta}_{t-1} + \boldsymbol{\varepsilon}_t$$

$$\mathbf{e}_t \sim \text{independent } N(\mathbf{0}, \boldsymbol{\Sigma}_t)$$

$$\boldsymbol{\varepsilon}_t \sim \text{independent } N(\mathbf{0}, \mathbf{Q}_t)$$

$$\hat{\mathbf{Y}}_{rt} \sim N\left(\mathbf{Z}_{rt}\hat{\boldsymbol{\beta}}_{t|t} , \boldsymbol{\Sigma}_{rrt} + \mathbf{Z}_{rt}\boldsymbol{\Omega}_{t|t}\mathbf{Z}'_{rt}\right)$$

$$\hat{\boldsymbol{\beta}}_{t|t} = \mathbf{M}\hat{\boldsymbol{\beta}}_{t-1|t-1} + \mathbf{K}_t\left(\mathbf{Y}_{ot} - \mathbf{Z}_{ot}\mathbf{M}\hat{\boldsymbol{\beta}}_{t-1|t-1}\right)$$

$$\boldsymbol{\Omega}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{Z}_{rt})\boldsymbol{\Omega}_{t|t-1}$$

$$\mathbf{K}_t = \boldsymbol{\Omega}_{t|t-1}\mathbf{Z}'_{rt}(\mathbf{Z}'_{rt}\boldsymbol{\Omega}_{t|t-1}\mathbf{Z}_{rt} + \boldsymbol{\Sigma}_{oot})^{-1}$$

$$\boldsymbol{\Omega}_{t|t-1} = \mathbf{M}\boldsymbol{\Omega}_{t-1|t-1}\mathbf{M}_t + \mathbf{Q}_t$$

Dynamic Logistic Regression Model

$$\mathbf{m}_{it} \sim \text{Ber}(\text{p}_{jt}), \mathbf{x}_{jt} \sim \text{Bin}(\text{c}_{jt}, \text{p}_{jt})$$

$$\text{p}_{jt} = \left[1 + exp\left(-\mathbf{Z}'_{jt}\boldsymbol{\gamma}_t\right)\right]^{-1}$$

$$\boldsymbol{\gamma}_t = \mathbf{H}\boldsymbol{\gamma}_{t-1} + \boldsymbol{\varepsilon}_t \quad , \quad \boldsymbol{\gamma}_t \perp \mathbf{Z}_t$$

$$\boldsymbol{\gamma}_1 \sim N\left(\boldsymbol{\gamma}_0, \Xi_{\boldsymbol{\gamma}_0}\right)$$

$$\boldsymbol{\varepsilon}_t \sim \text{independent } N\left(\mathbf{0}, \Xi_t\right)$$

$$\Pr(\hat{\mathbf{m}}_{it} = 1) \cong \left[1 + exp\left(-\mathbf{Z}'_{jt}\hat{\boldsymbol{\gamma}}_{t|t}\right)\right]^{-1} = \hat{p}_{it}$$

$$Var(\mathbf{m}_{it} = 1) = \hat{p}_{it}(1 - \hat{p}_{it})$$

$$\hat{\boldsymbol{\gamma}}_{t|t} = \mathbf{H}\hat{\boldsymbol{\gamma}}_{t-1|t-1} + \Sigma_{t|t-1}\mathbf{Z}'_{ot}\{\mathbf{x}_{ot} - \mathbf{C}_t\hat{p}_t\}$$

$$\Sigma_{t|t} = (\mathbf{I} - \mathbf{G}_t)\Sigma_{t|t-1}$$

$$\mathbf{G}_t = \Sigma_{t|t-1}\left(\mathbf{Z}'_{ot}\,diag(\hat{p}_{jt}(1 - \hat{p}_{jt}))\mathbf{Z}_{ot} + \Sigma_{t|t-1}\right)^{-1}$$

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

Australian Bureau of Statistics

Concluding remarks

- These methods apply to a large number of Big Data applications
  - For example, mobile phone data
  - The challenge (cost, feasibility etc.) is availability of ground truths/measurements
- For official statistics, there is a role for survey sampling even with Big Data

References:

Tam, S & Clarke, F 2015 (2015). 'Big Data, Official Statistics and Some Initiatives of the ABS', *International Statistical Review,* to appear.

Tam, S (2015). *A Statistical Framework for Analysing Big Data*, *June 2015,* cat. no. 1351.0.55.056, ABS, Canberra.

siu-ming.tam@abs.gov.au