-

Satellite Imagery and an ABS Methodology for Predicting Crop Yields

Dr Siu-Ming Tam
Chief Methodologist
Global WG on Big Data
Beijing, China
October, 2014

Australian Bureau of Statistics

Outline

Caveats

I. Expert?

II. Methodology

Two parts of the talk

I. Satellite Imagery basics

II. An ABS application

Part I

Some views on Big Data

Satellite imagery basics

Challenges

Partners

Types of satellites

A satellite is an object that moves around a larger object

- Earth around the sun

Human made satellites

- Revolves around the earth to collect info and communicates back to earth
- About 3,000 operating in earth orbit

Source: Sam Batzil, WisconsinView.org


GOES-R. Credit: NOAA


Landsat-7. Credit: NASA


Sentinel Ib. Credit: European Space Agency


QuickSCAT. Credit: NASA


GPS Block IIIA. Credit: Lockheed Martin


NOAA-18. Credit: NOAA

Types of satellites

Weather and atmosphere monitoring (e.g. GOES_R)

Earth observation and mapping (e.g. Landsat7)

Astronomical and Planetary Exploration

Communication

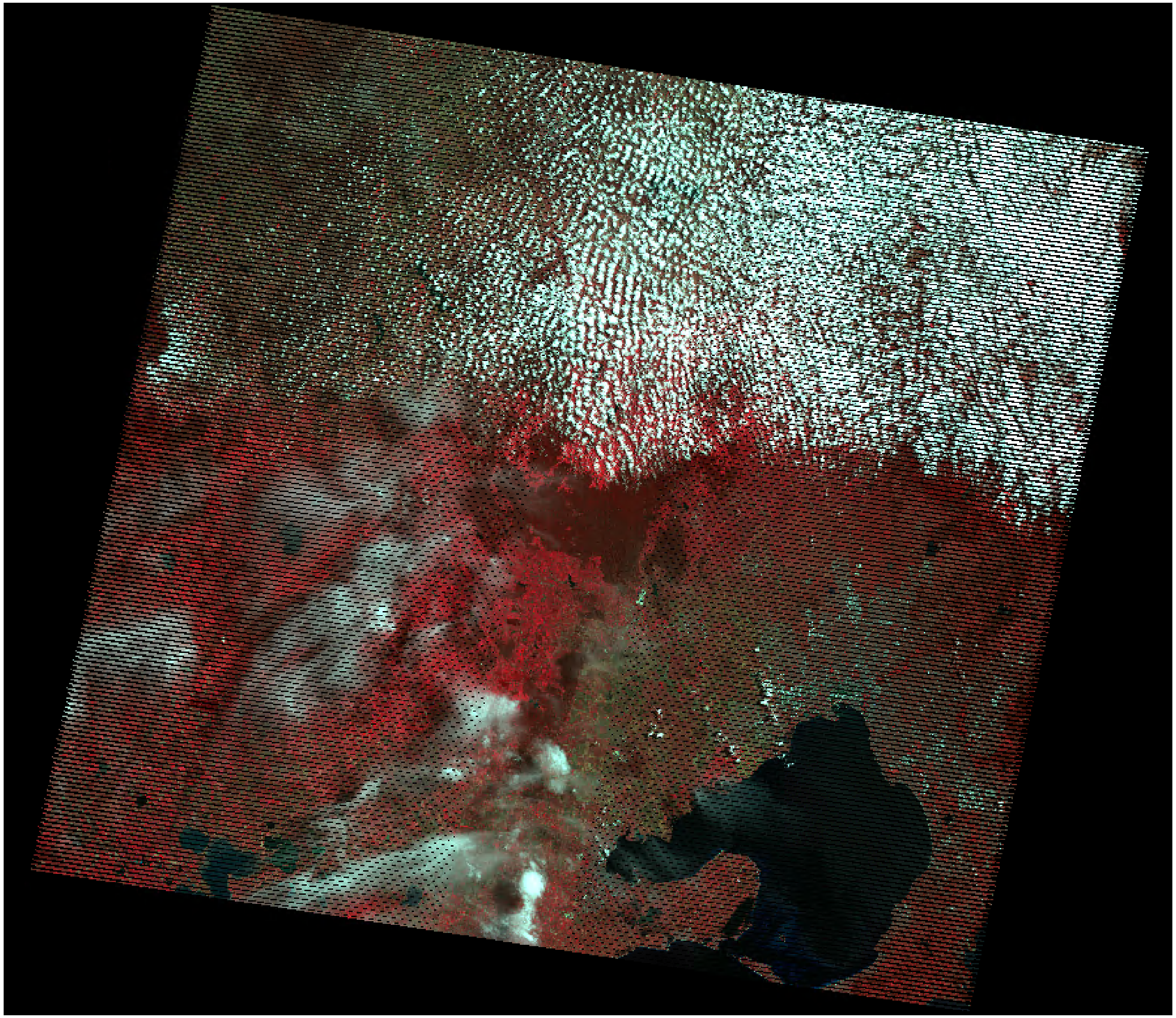
Navigation (GPS)

Military

Sensors are instruments that record solar, radar or laser radiation signals from reflection of earth objects.

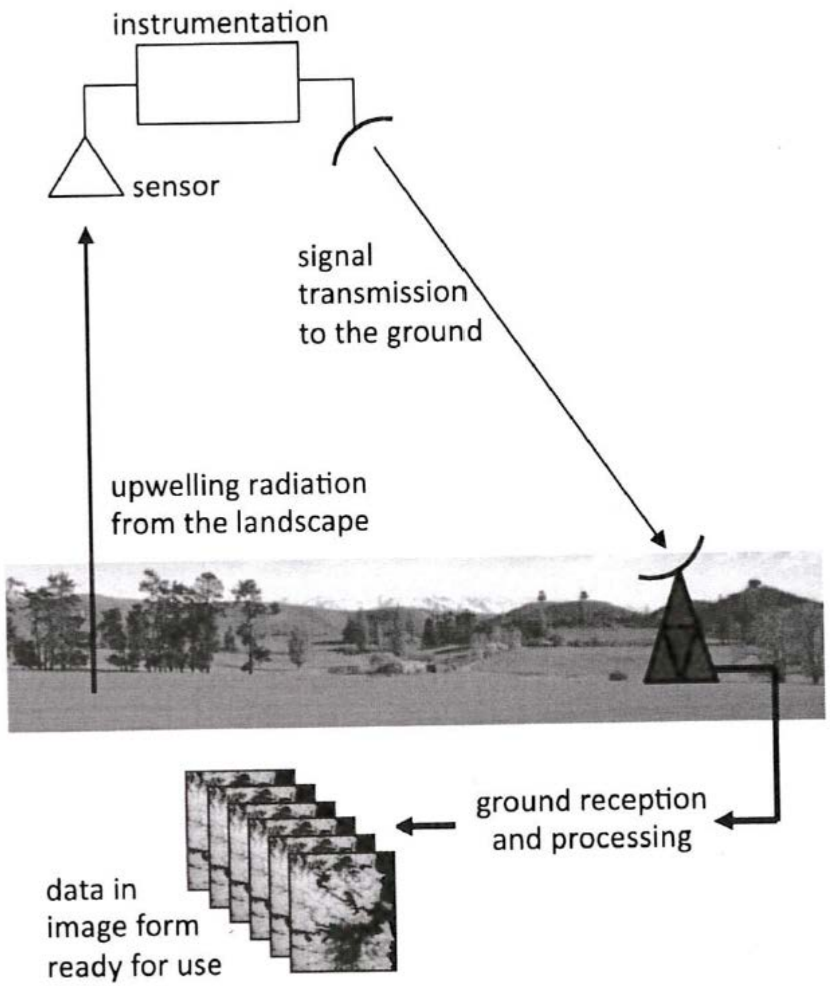
Source: Sam Batzil, WisconsinView.org

Satellite imagery data basics

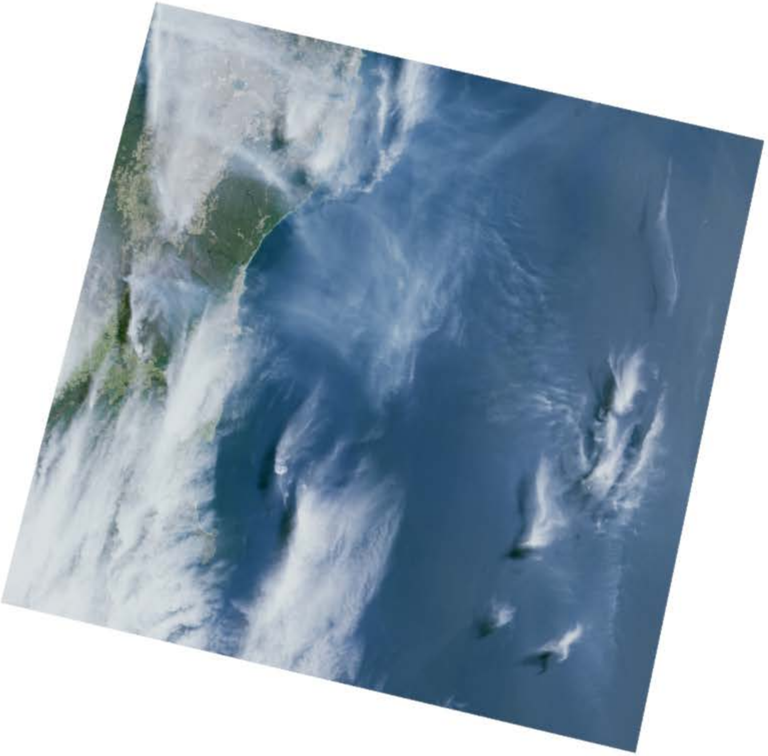


Satellite imagery data basics

Australian Bureau of Statistics

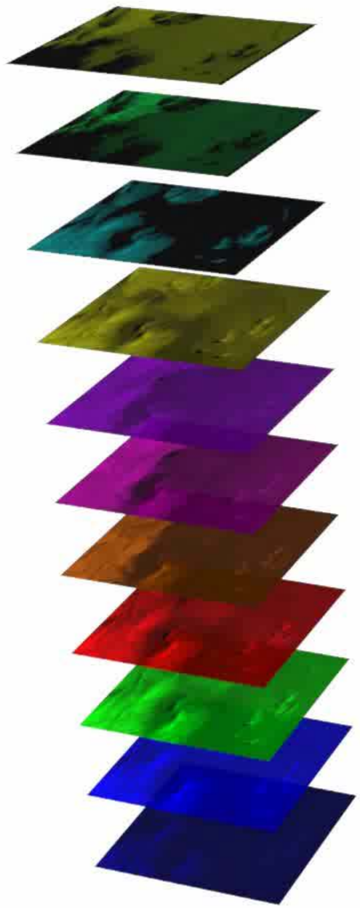


Satellite Images: Not just ‘photos’



Satellite Image

Not just 'photos'



11.50-12.51 - TIRS 2
10.60-11.19 - TIRS 1
1.36-1.38 - Cirrus
0.5 - 0.68 - Panchromatic
2.11-2.29 - SWIR 2
1.11-2.29 - SWIR 1
0.85-0.88 - Near Infrared
0.64-0.67 - Red
0.53-0.59 - Green
0.45-0.51 - Blue
0.43-0.45 - New Deep Blue

Satellite imagery data 101

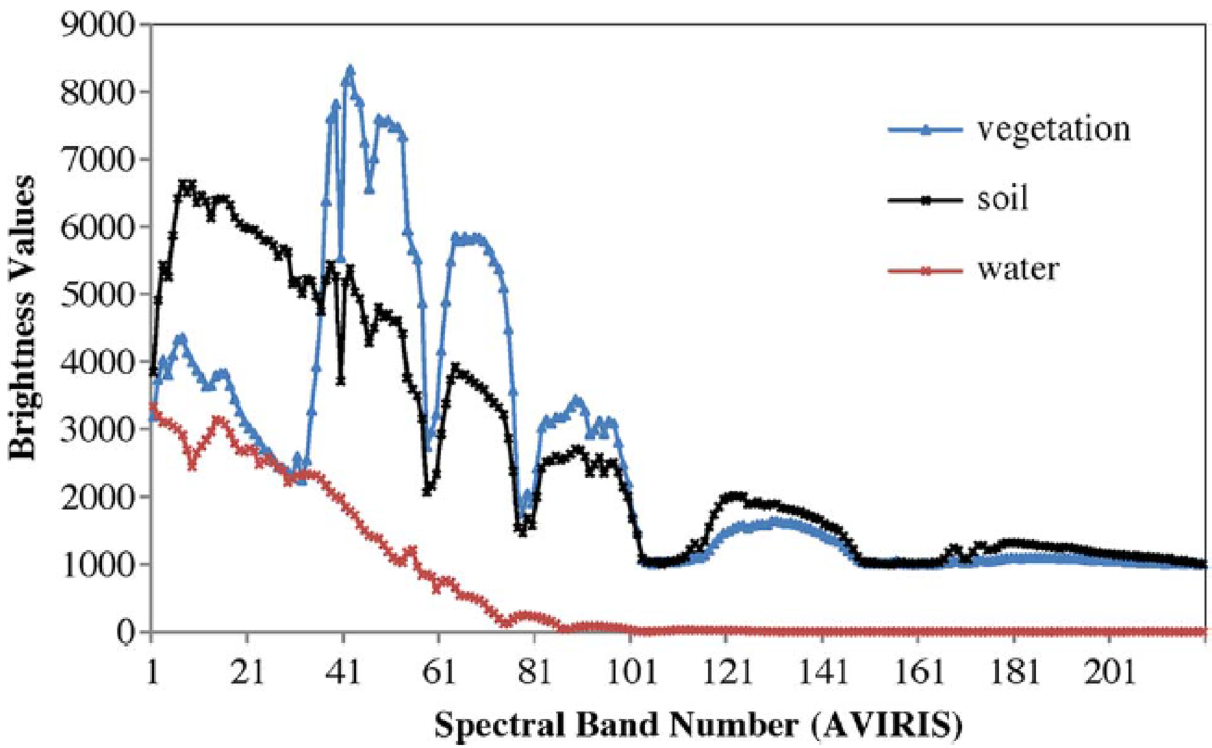


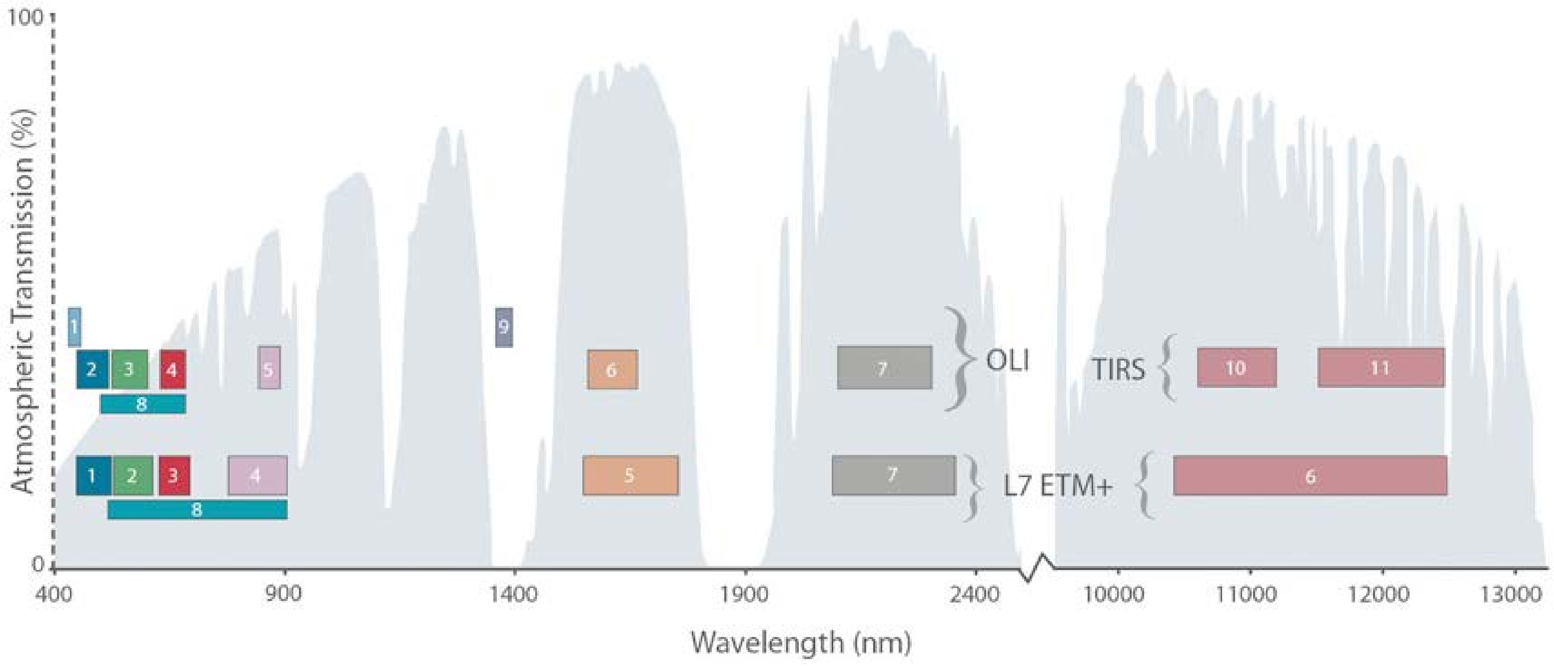
Fig. 1. ***Examples of vegetation, soil, and water spectra recorded by AVIRIS.***

Band designations for LandSat 7 & LandSat 8

Australian Bureau of Statistics

Multispectral data rather than hyperspectral data

We are currently using LandSat 7 data



Landsat 7 data

Landsat 7 launched in April, 1999 to refresh satellite photos of the world

Imagery available once every 16 days per pixel (25m * 25m) covering the globe

Each pixel has 7 reflectance (or radiance)

Images may be downloaded free of charge from US Geological Survey (http://earthexplorer.usgs.gov/)

In May, 2003 Scan Line Corrector failure led to 22% of the data missing

Landsat 7 was joined by Landsat 8 in 2013.

Large manual process to match farm location with pixels for ground truth data (to create a training dataset)

Experimental analyses by ABS only downloaded a small dataset into our Big Data Laboratory so no issue about storage

Australian Bureau of Statistics

Landsat data are organised as separate "collections"; so huge manual process to create a "time series" of pixels. Also data were not corrected for movement of the continent.

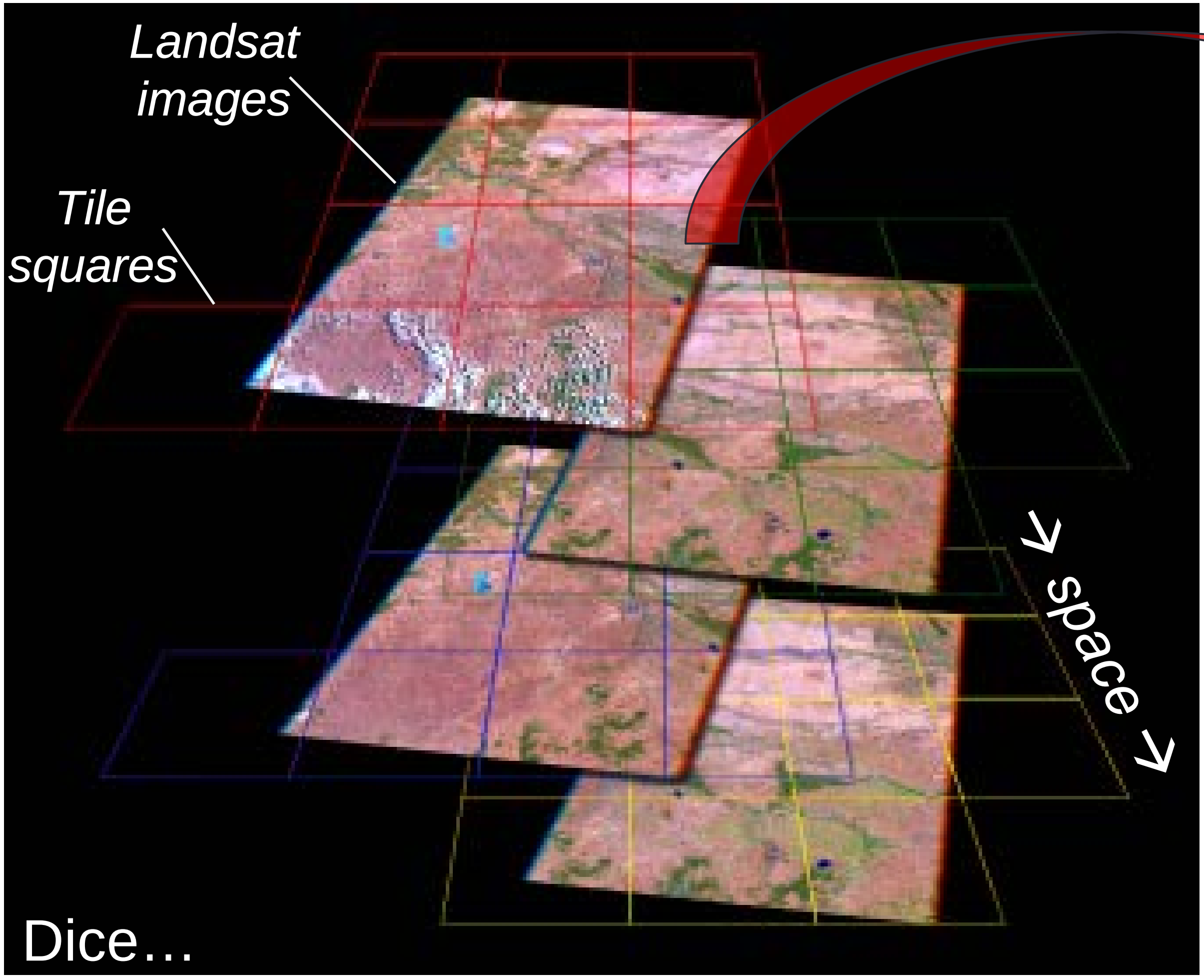
- We intend to "interrogate" the Australian Data Cube for future analysis

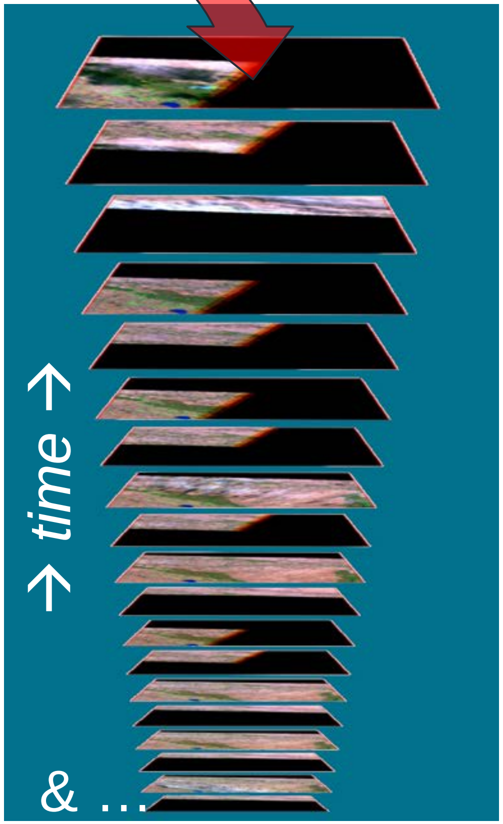
Methodology

- ABS has recently developing a methodology for predicting crop yields using ground truth data – to be discussed in Part II
- The algorithms have yet to be tested with large amount of ground truth data
- Still talking to various possible providers of ground truth data

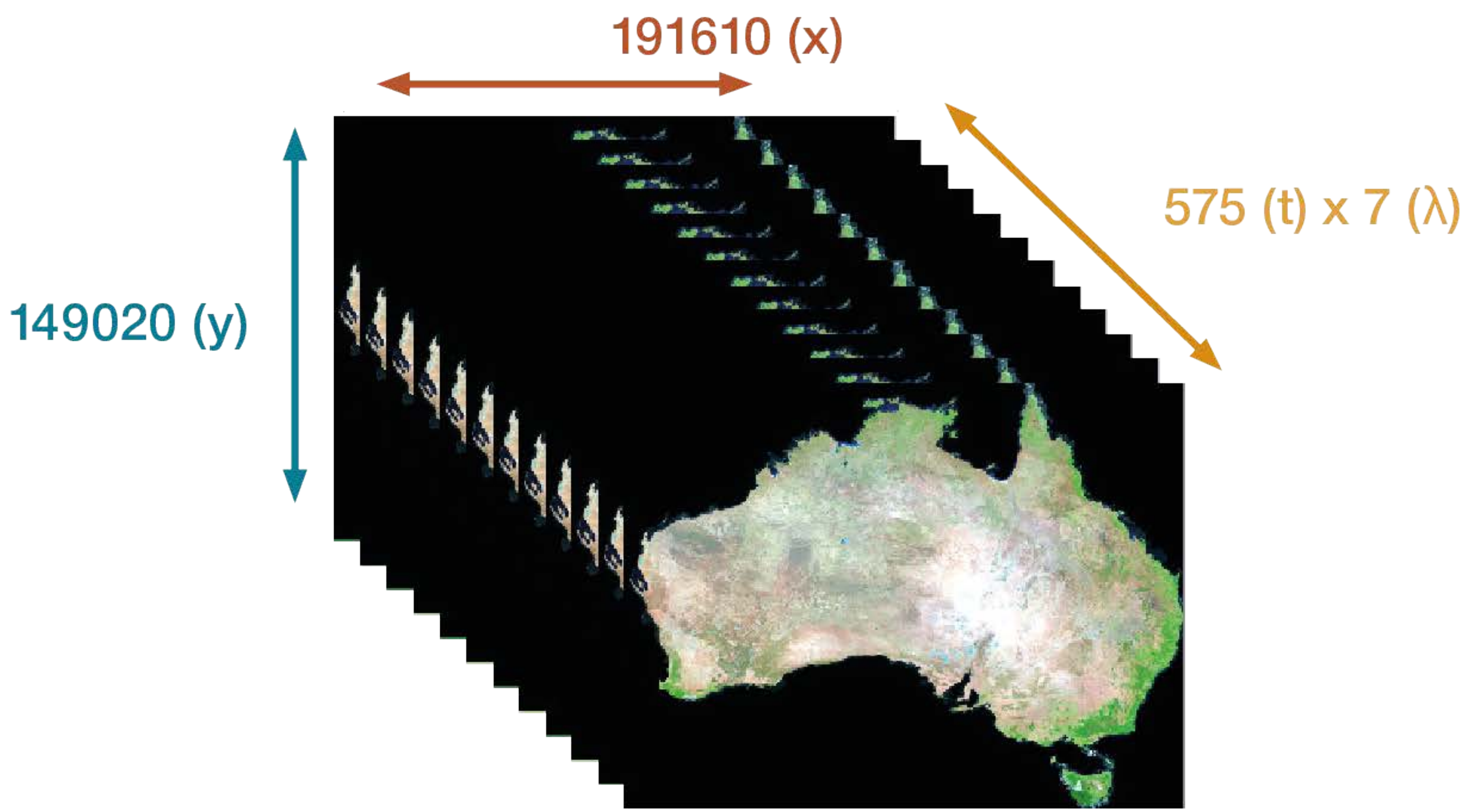
A better source for satellite imagery data for Australia is becoming available – The Australian Data Cube

"Cubing" Landsat images





The Australian Data Cube concept



Data cube

Created by Bureau of Metrology, CSIRO, National Computing Infrastructure and Geoscience, Australia using Landsat Satellites

- about 4.5 PB
- Library of Congress Books is 10 TB; 1 PB = 1000 TB

Data continued to be prospectively and standardised into a common framework

- So analysts can concentrate on analysis, rather than data assembly

Analysts can 'drill down' all the data about a particular location, at a pixel level, and access all historical yet comparable Landsat data

GA intends to load European and Japanese satellite data from 2015

- Satellite imagery available every 10 minutes!
- I TB of data per day

Data loaded in the National Computing Infrastructure which houses the high performance computers

- ABS does not need to store this data
- ABS can use the virtual computing environment to “play” with the data

Thanks to Dr S Minchin of GA for providing the Data Cube slides

Challenges

Australian Bureau of Statistics

What problems satellite imagery data are going to solve?

Business case for it

- Efficiency?
- More frequently or timely data?
- Data at a small area level?
- Prediction/forecasting? Is this our core business?
- Replacing/complementing official data?
- Cost benefit?

Methodology for analysing the data

- Handling missing data e.g. Landsat 7 sensor problems
- Handling missing data from cloud covers
- Scientific modelling vs statistical modelling
- Algorithms

Sourcing the data

. Satellite imagery

. Ground truth data – if statistical modelling is to be adopted

Maintaining trust of official statistics

- Quality assessments

Partners

Direction setting groups

- HLG – Modernisation of Statistics

Partners in methodology

- Research organisations
- Academics

Providers of data

- Satellite imagery providers
- Ground truthers

User of new official statistics

Management and staff

Other NSOs

- National Bureau of Statistics, China
- National Agricultural Statistics Service (US)
- Statistics Netherlands
- Statistics Canada
- INEGI, Mexico
- Dane, Colombia
- Others?

In Australia

Australian Bureau of Statistics

Research, collection and archiving effort carried out by:

- Geoscience Australia
- BoM
- ABARES
- ACEMS
- CSIRO
- Curtin University
- ADFA & UNSW
- TERN
- Sense-T
- Landgate – Satellite Remote Sensing Services
- WASTAC

Questions?

Siu-Ming.Tam@abs.gov.au

Part II – ABS Example

ABS Big Data Strategy

ABS Flagships

What problem we are trying to solve?

What methodology to use for analysis?

What is our research problem?

Rather than exclusively through a

***traditional survey** collection,*

is it possible to use

satellite imagery data

to estimate the

area of land used to grow different crops

and crop yields

in Australia?

Why?

Potential to reduce costs by

- Reducing the sample size for Agricultural surveys

Provision of more frequent data

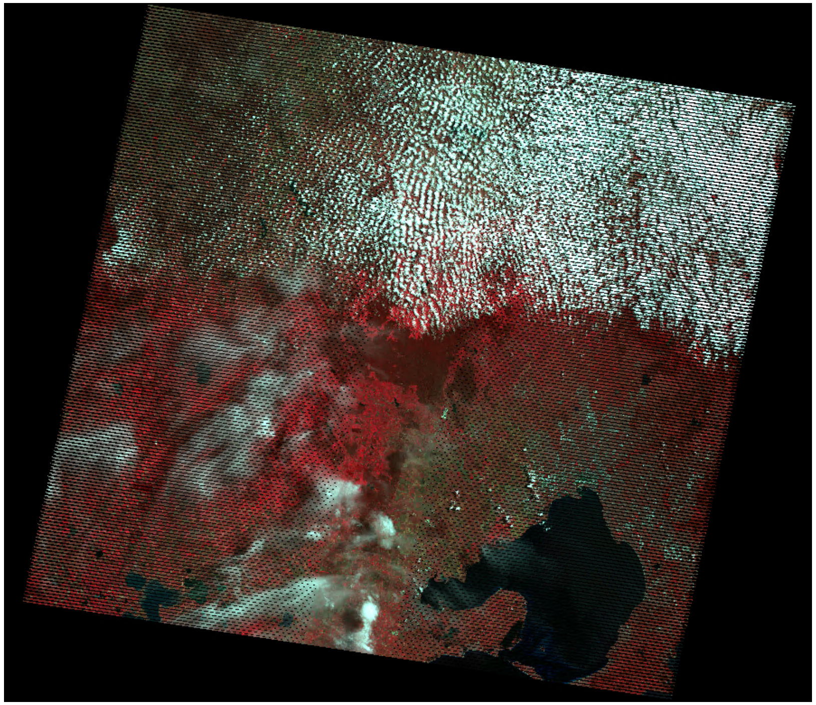
Provision of small area data

Business case has yet to be established

- Current priority is to test the efficacy of the methodology

Estimating crop yields from

Satellite imagery



The data

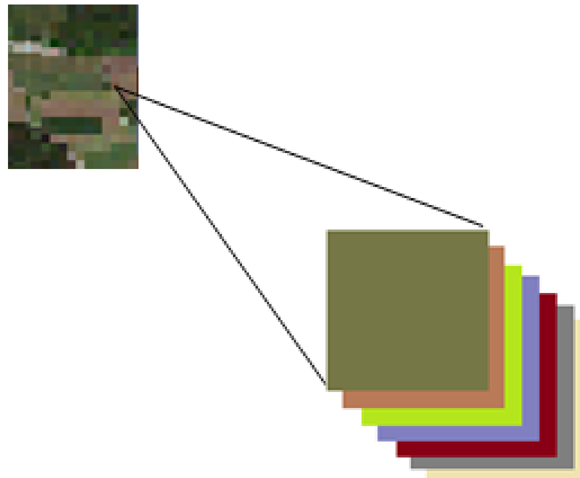
Landsat 7 imagery from US Geological Survey

- reflectance data from 7 freq bands for pixels of 25x25 m²

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

Pixel classification and yields

Australian Bureau of Statistics



7 reflectance measurements
$\boldsymbol{y} = (y_1, \ldots, y_7)$

$f(\boldsymbol{y}) = c$

Big Data = Big Traps?

Two broad types of errors in sampled data sets

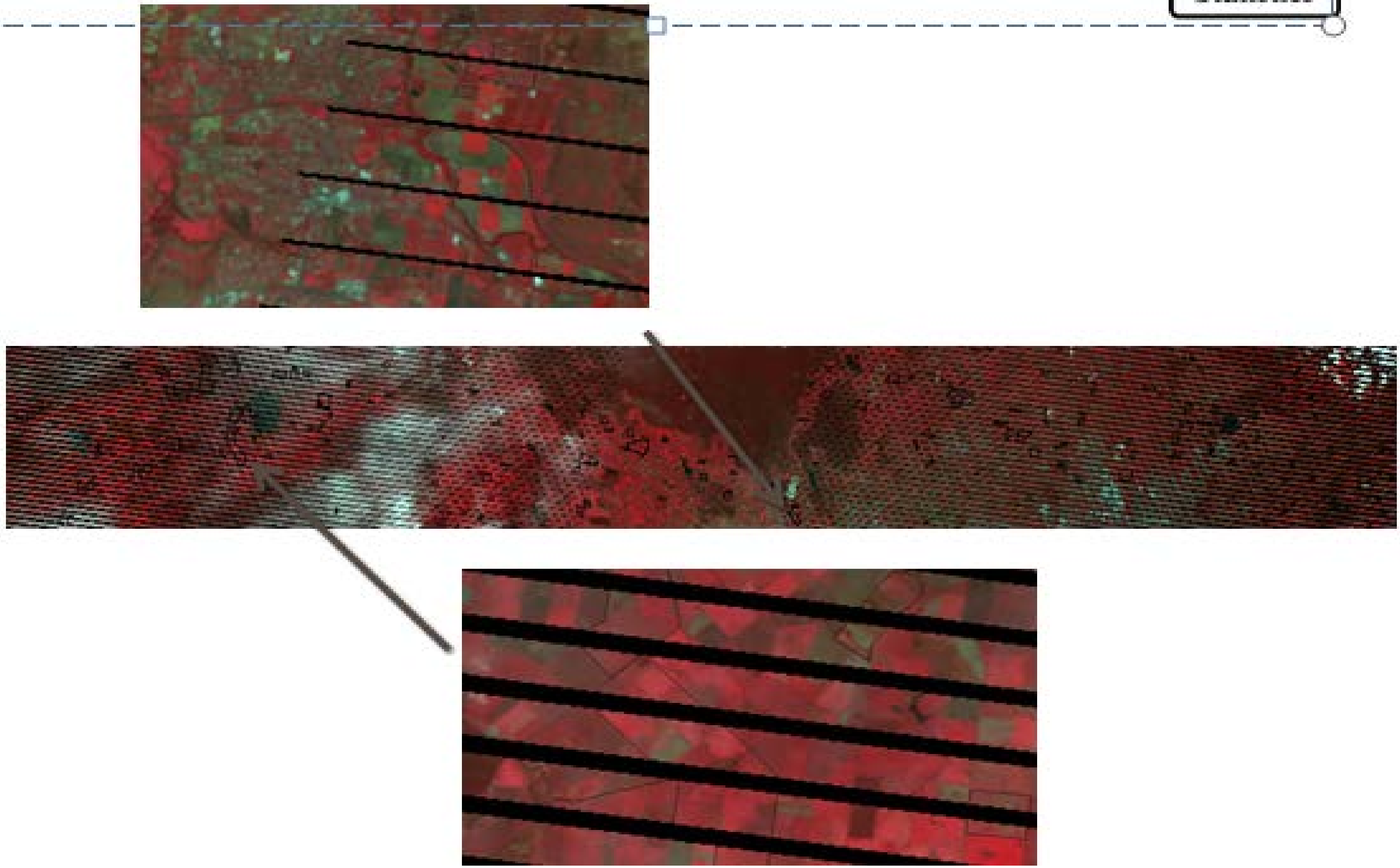
- Sampling error
  - Dependent on size
- Non sampling error
  - Coverage bias - Big Data population is not the population
  - Self selection bias – squeaky wheels
  - Representation bias – multiple representation
  - Measurement error
  - Increasing the sample size does NOT reduce non-sampling errors

Traps

- Big Data is a solution is search of a problem
- Putting the cart before the horse
- Correlation = causality

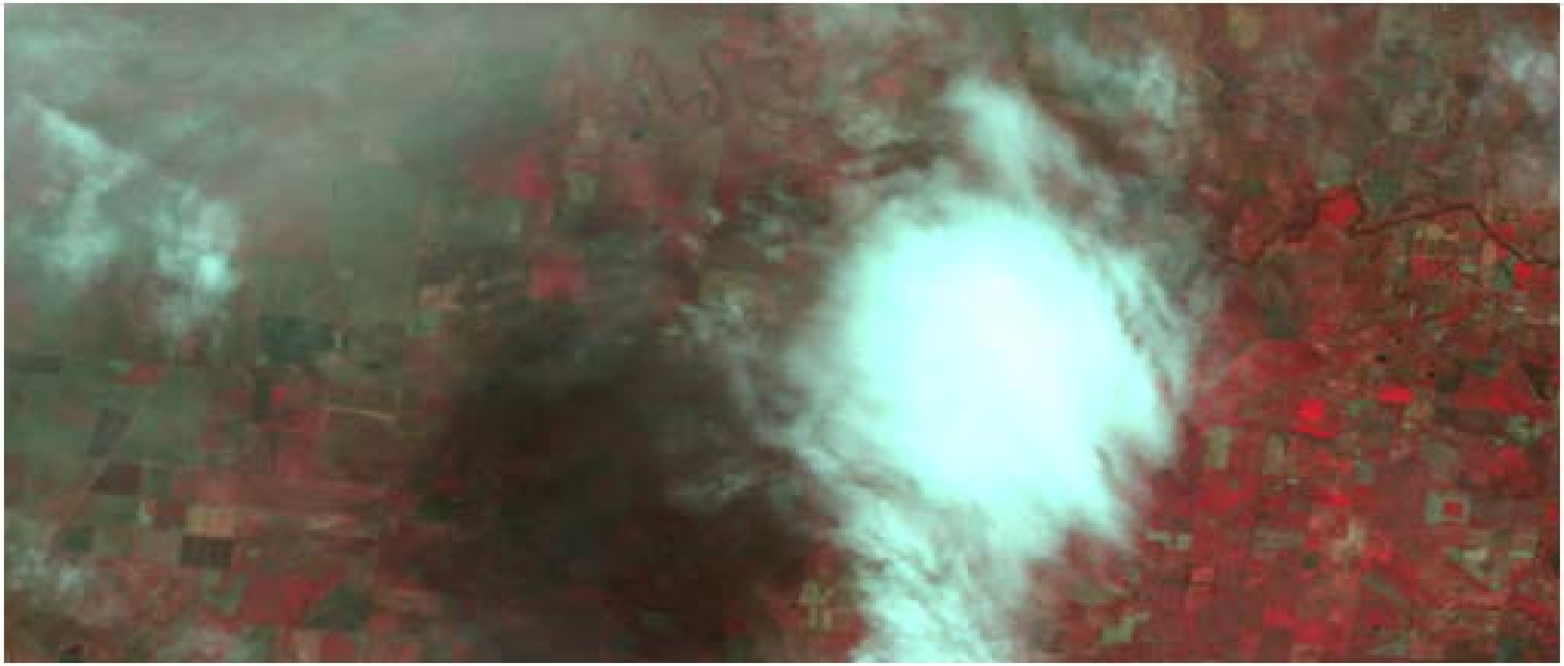
Missingness from equipment problems

Australian Bureau of Statistics

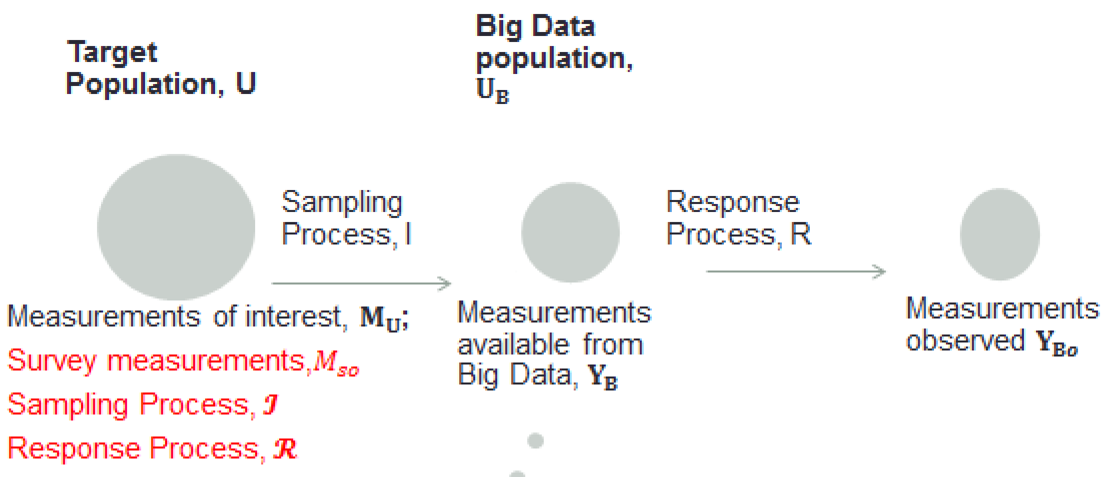


Perpetual cloud cover

Missing not at random



Survey (or Design) Data and Big (or Organic) Data



Transformation Model - $f(\mathbf{M_U} | \mathbf{Y_U}; \boldsymbol{\varphi})$

Process model - $f(\mathbf{Y_U}; \boldsymbol{\theta})$

Parameter model - $f(\boldsymbol{\varphi})$ and $f(\boldsymbol{\theta})$

Data - $\mathbf{M_{so}}, \mathbf{Y_{Bo}}$; Processes - $\mathcal{I}, \mathcal{R}$, I, R; time dimension as well

Bayesian Inference Framework

Predictive (correct) inference for $\mathbf{M_U}$is:

The conditional probability density function (CPDF) of $\mathbf{M_U}$ given $\mathbf{M_{so}}, \mathbf{Y_{Bo}}, \mathcal{I}, \mathcal{R}$, I, R.

Generally there is no closed form for this function.

However, under certain conditions – see next slide – the CPDF is the same as

- the CDPF of $\mathbf{M_U}$ **given** $\mathbf{M_{so}}, \mathbf{Y_{Bo}}$, I, R, i.e. the missingness due to sampling can be ignored; and
- The CDPF of $\mathbf{M_U}$ **given** $\mathbf{M_{so}}, \mathbf{Y_{Bo}}$ i.e. the missingness due to Big Data membership can be ignored

Bayesian Inference Framework

Predictive (correct) inference for $\mathbf{M_U}$ is:

$$f(\mathbf{M_U}|\mathbf{M_{so}}, \mathbf{Y_{Bo}}, \mathcal{I}, \mathcal{R}, \mathrm{I}, \mathrm{R}) \propto \iiint f(\mathbf{M_U}, \mathbf{M_{so}}, \mathbf{Y_{Bo}}, \mathbf{Y_C}, \mathcal{I}, \mathcal{R}, \mathrm{I}, \mathrm{R}, \boldsymbol{\theta}, \boldsymbol{\varphi}) d\boldsymbol{\theta} d\boldsymbol{\varphi} d\mathbf{Y_C}$$

(generally no closed form) where $\mathbf{Y_{Bo}} \cup \mathbf{Y_C} = \mathbf{Y_U}$, or

$$f(\mathbf{M_U}|\ \mathbf{M_{so}}, \mathbf{Y_{Bo}}, \mathrm{I}, \mathrm{R})$$

provided that

$$f(\mathcal{I}, \mathcal{R}|\mathbf{M_U}, \mathbf{M_{so}}, \mathbf{Y_{Bo}}, \mathrm{I}, \mathrm{R}) = f(\mathcal{I}, \mathcal{R}|\mathbf{M_{so}}, \mathbf{Y_{Bo}}, \mathrm{I}, \mathrm{R}) \text{ (controlled by sampler)},$$

or

$$f(\mathbf{M_U}|\ \mathbf{M_{so}}, \mathbf{Y_{Bo}})$$

provided that

$$f(\mathrm{R}, \mathrm{I}|\mathbf{M_U}, \mathbf{Y_{Bo}}, \mathbf{Y_C}, \boldsymbol{\theta}, \boldsymbol{\varphi}) = f(\mathrm{R}, \mathrm{I}|\mathbf{M_U}, \mathbf{Y_{Bo}}) \text{ (controlled by BD participants).}$$

Missingness

In English

- The missing process for the survey sample can be ignored if missingness does not depend on the probability of growing a targeted crop
  - Easy to fulfil as the sampling process is determined by the official statistician
- The missing process for Big Data (BD) can be ignored if missingness does not depend on the observations from BD
  - Hard to control as participation in some BD platforms is voluntary and by self selection.
- Modelling may be required in other situations
  - Modelling is hard work
  - Computation is hard, as there is generally no closed form solution

Predicting crop yields – Methodology in English – assuming Missing At Random

for every pixel:

I. Yield (Y) = Crop type (m) * quantity (q) (or ln Y = ln m + ln q)

II. Assume m follows a logistic regression model, but allowing the regression coefficients to change over time

   1. To allow for different electromagnetic spectra emitted from maturing crops

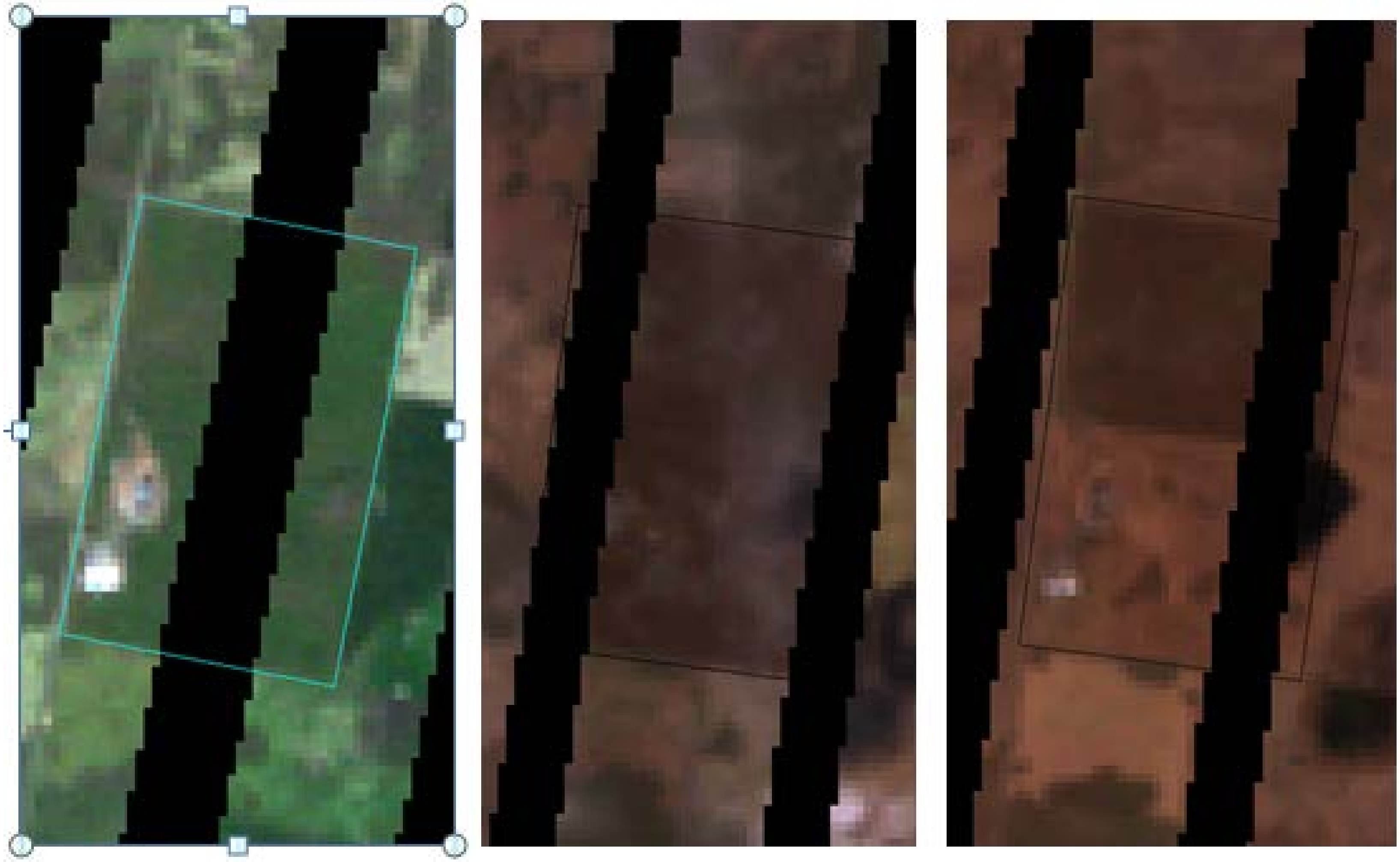
   2. Independent variables are reflectance

III. Assume ln q follows a logistic normal regression model, also allowing the regression coefficients to change over time

   1. Independent variables are land surface temperature and moisture

Modelling for variation over time

9 Nov 2010 | 4 Dec 2010 | 5 Jan



How to predict crop areas

Australian Bureau of Statistics

$\mathrm{m}_{\mathrm{ti}} = (1 + e^{-\mathbf{Y}_{\mathrm{ti}}'\boldsymbol{\beta}_{\mathrm{t}}})^{-1}$

$\boldsymbol{\beta}_{\mathbf{t}} = \boldsymbol{\beta}_{\mathbf{t-1}} + \boldsymbol{\varepsilon}_{\mathbf{t}}$, $\boldsymbol{\beta}_{\mathbf{t}} \perp\!\!\!\perp \mathbf{Y}_{\mathbf{t}}$,

$\boldsymbol{\varepsilon}_{\mathbf{t}} \sim$ independent $\mathbf{N}(0, \Omega_{\mathbf{t}})$, $\boldsymbol{\varepsilon}_{\mathbf{t}} \perp\!\!\!\perp \mathbf{D}^{(\mathbf{t})}$,

Step A - At time t, select a random sample of pixels as a “training data set”

Step B - For each pixel, use the Landsat data to obtain the 7 reflectance

Step C - For the same pixel, seek “ground truths”, i.e. undertake field work to find out whether the pixel is growing the targeted crop or not (Yes = 1, and No = 0)

Step D - Stack these data up to form $\mathbf{Y}_{\mathbf{ts}}$, and $\mathbf{M}_{\mathbf{ts}}$

Step E - Use Newton-Raphson algorithm to calculate $\widehat{\boldsymbol{\beta}}_{\mathbf{t|t}}$ from

$\widehat{\boldsymbol{\beta}}_{\mathbf{t|t}} = \widehat{\boldsymbol{\beta}}_{\mathbf{t-1|t-1}} + \sum_{\mathbf{t|t-1}}^{-1}\{\mathbf{Y}_{\mathbf{ts}}'\mathbf{M}_{\mathbf{ts}} - \mathbf{Y}_{\mathbf{ts}}'\, \boldsymbol{\sigma}(\mathbf{Y}_{\mathbf{t}}'\widehat{\boldsymbol{\beta}}_{\mathbf{t|t}})\}$ **– see Theorem 1**

How to predict crop quantities

$\mathbf{M_{tB}}|\mathbf{Y_{tB}},\boldsymbol{\beta}_{\mathbf{t}} \sim \mathrm{N}(\mathbf{Y_{tB}}\boldsymbol{\beta}_{\mathbf{t}},\Sigma_{\mathbf{t}})$ where $\mathbf{M_{tB}} = \ln \mathbf{Q_{ts}}$

i.e. $\mathrm{E}(\mathrm{m_{ti}}|\ \mathbf{Y_{ti}}, \boldsymbol{\beta}_{\mathbf{t}}) = \mathbf{Y_{ti}}'\boldsymbol{\beta}_{\mathbf{t}}$

$\boldsymbol{\beta}_{\mathbf{t}} = \boldsymbol{\beta}_{\mathbf{t-1}} + \boldsymbol{\varepsilon}_{\mathbf{t}}$, $\boldsymbol{\beta}_{\mathbf{t}} \perp\!\!\!\perp \mathbf{Y_t}$

$\boldsymbol{\varepsilon}_{\mathbf{t}} \sim$ independent $\mathbf{N}(0, \Omega_{\mathbf{t}})$, $\boldsymbol{\varepsilon}_{\mathbf{t}} \perp\!\!\!\perp \mathbf{D}^{(\mathbf{t})}$

Step A - At time t, for each of the sample of pixels selected to predict probabilities:

- seek "ground truths", i.e. undertake field work to find out the quantities of crop produced; and
- Obtain values of the covariates ie LST, moisture from weather satellites, $\mathbf{y_{ti}}$

Step B - Stack these data up to form $\mathbf{Y_{ts}}$, and $\mathbf{M_{ts}}$ (= $\ln\mathbf{Q_{ts}}$)

Step C $-$ Calculate $\widehat{\boldsymbol{\beta}}_{\mathbf{t|t}}$ from

$\widehat{\boldsymbol{\beta}}_{\mathbf{t|t}} = \widehat{\boldsymbol{\beta}}_{\mathbf{t-1|t-1}} + \Sigma_{\mathbf{t|t}}\ \mathbf{Y}'_{\mathbf{ts}}\ \Sigma^{-1}_{\mathbf{tss}}(\mathbf{M_{ts}} - \mathbf{Y}'_{\mathbf{ts}}\widehat{\boldsymbol{\beta}}_{\mathbf{t-1|t-1}})$ – **see Theorem 2**

Take home messages

Business case for Big Data

Methodology to provide valid statistical inference for Big Data

- Combining survey with Big Data
- Business case from reduction in survey sample sizes

Model for predicting crop yields (applies to all counts and continuous data)

- Algorithms developed
- Cross validation of algorithms required ground truth data
- Model ignored missing data – not a major problem for satellite imagery data, but will be for e.g. social media data

Key References

Tam, S.M. and Clark, Frederic (2014) Big Data, Official Statisticis and Some Initiatives of the Australian Bureau of Statistics. Paper submitted for publication

Johnson, David 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, 116-128

Questions?

Siu-Ming.Tam@abs.gov.au

Panel Discussion Questions

Australian Bureau of Statistics

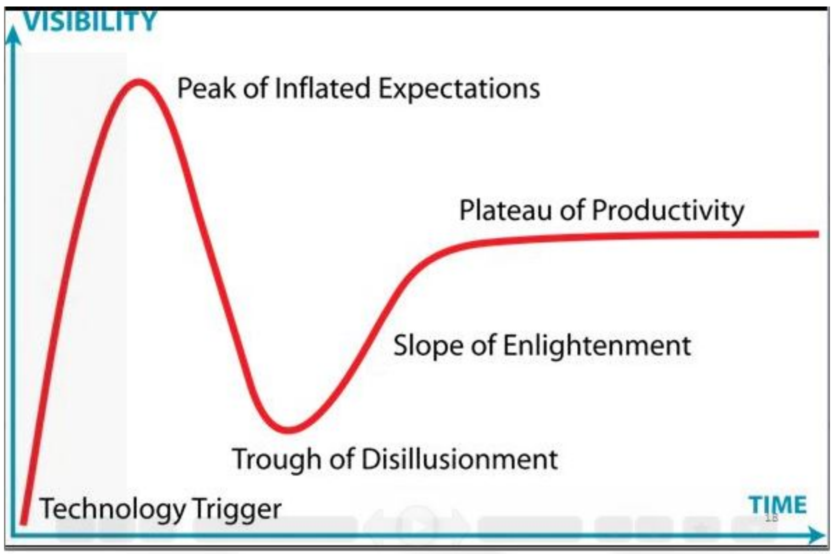
How will Big Data benefit your institute?

Could it benefit developing countries as well?

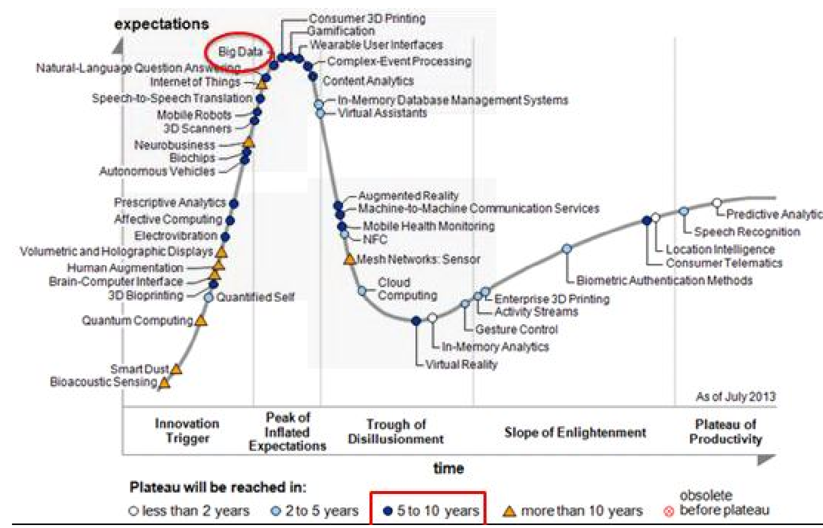
Will Big Data help in getting timelier and more indicators for the Post-2015 development agenda?

Big Data = Big Hype?

Gartner Hype Curve



Big Data on the Hype Curve



ABS Big Data Strategy

Australian Bureau of

ABS Capability

- Authority for data acquisition
- Authorised Integrator of sensitive data
- Ability to integrate with Census and Survey data
- Trust in the ABS and our reputation for Integrity, Impartiality and Quality

Our Objective:

Effective application of big data to reduce costs, improve timeliness, quality, and expand the range of our statistics.

- Identify statistical needs that should be the focus of early efforts to apply big data
- Identify "high potential" data sources
- Seek funding and support for the application of big data
- Undertake pilot applications to better understand the barriers, enablers and value proposition

Needs

- Population movements
- Environment
- Prices

Sources

- Satellite
- Telecom
- Financial Sector
- Retail Prices
- Utilities

Research Partners

- Big Data Research Partnerships
- ARC Partner Investigator
- APS Big Data Working Group & Analytics COE
- UNECE Big Data Working Group

Key Enabler:

Active partnership and collaboration with those who can help us apply big data

- Government Agencies
- Academics and Researchers
- Private custodians of big data
- Working Groups and Centres of Excellence

Key Enabler:

Enhanced ABS capability to use big data

- Develop the skills of our staff
- Establish the infrastructure needed to exploit big data
- Develop appropriate methods and techniques

ABS Big Data Research areas - Flagship

Satellite imagery data for agricultural statistics

Multiply-linked employer-employee data for productivity analysis

Mobile positioning data for measuring population mobility

Predictive modelling of survey non-response behaviour

Data visualisation techniques for exploring large datasets

Predictive modelling of unemployment for small areas

(in decreasing order of progress of development)

Big Data and Big Opportunities

Possible benefits	Statistical activities
• Replace direct data collection	Sample frames or registers
• Complementary direct data collection	Small domain estimation
• Substitute data items	Small population group estimation
• New data items	Enabling data imputation, editing and confrontation
• Supplementary information to improve quality	Enabling data linking and fusion
	Producing new statistical products
	Improving statistical operations

Australian Bureau of Statistics

Big Data and Big Challenges

ABS objective

Harness Big Data sources to to create a richer, more dynamic and focused statistical picture of Australia for better informed decision-making

Challenges

Business benefit

Privacy and public trust

Technological feasibility

Data acquisition

Data integrity

Methodological soundness

- How to make valid statistical inferences

Big Data = Big Traps?

Two broad types of errors in sampled data sets

- Sampling error
  - Dependent on size
- Non sampling error
  - Coverage bias - Big Data population is not the population
  - Self selection bias – squeaky wheels
  - Representation bias – multiple representation
  - Measurement error
  - Increasing the sample size does NOT reduce non-sampling errors

Traps

- Big Data is a solution is search of a problem
- Putting the cart before the horse
- Correlation = causality

Australian Bureau of Statistics

Big Data = Big Sources, but

not entirely foreign to official statisticians

Eg Administrative records, Scanner Data

- **Administrative Records**
  - PIT Records
  - Medical Records
  - Bank Records
- **Commercial Transactions**
  - Credit Card Transactions
  - Scanner Transactions
  - Online Purchases
- **Sensor Data**
  - Satellite Imagery
  - Ground Sensor Data
  - Location Data
- **Behaviour Metrics**
  - Search Engine Queries
  - Web Pages Views and Navigation
  - Media Subscriptions
- **Online Opinion**
  - Social Media Comments
  - Twitter Feeds

How will Big Data benefit ABS?

Australian Bureau of Statistics

Still an open question as we have yet to develop the business case for certain types of Big Data… But promising for

- Satellite Imagery Data
- Mobile phone data
- Harness own operational data

Could it benefit developing countries as well?

- Yes
  - Provided that sources are also available to DCs;
  - Methodology is available to them as well

Will Big Data help in getting timelier and more indicators for the Post-2015 development agenda?