1352.0.55.144

Australian Bureau of Statistics

Research Paper

Methodological Approaches for Utilising Satellite Imagery to Estimate Official Crop Area Statistics

New Issue

Research Paper

Methodological Approaches for Utilising Satellite Imagery to Estimate Official Crop Area Statistics

Jennifer Marley, Daniel Elazar and Kate Traeger

Analytical Services Branch

Methodology Advisory Committee

5–6 June 2014, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) FRI 19 SEP 2014

ABS Catalogue no. 1352.0.55.144

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 Mr Daniel Elazar, Analytical Services Branch on Canberra (02) 6252 6962 or email <analytical.services@abs.gov.au>.

METHODOLOGICAL APPROACHES FOR UTILISING SATELLITE IMAGERY TO ESTIMATE OFFICIAL CROP AREA STATISTICS

Jennifer Marley, Daniel Elazar and Kate Traeger
Analytical Services Branch

QUESTIONS FOR THE COMMITTEE

1. Is the Committee aware of any literature concerned with or have ideas about how to approach putting machine learning techniques, in particular support vector machines, on a firm statistical inferential basis?
2. Can the Committee provide advice on or suggest other statistical classifiers that perform well, in terms of both prediction accuracy and computational efficiency, on high-dimensional data that are likely to belong to similar and overlapping class distributions?
3. Is the Committee aware of time series techniques that perform well over short time spans (for example, the duration of a crop's growth from seeding to harvesting) that can be used to characterise and distinguish different crop types?
4. Does the Committee have advice regarding the implementation of spatial models to impute for missing values (either induced by systemic failures in the satellite sensors or cloud cover) and of measurement error models to utilise pixels affected by thin cloud cover?
5. Does the Committee think it is worthwhile pursuing agronomic auxiliary data that could be used to inform regional classifiers about the distributions of particular crop cultivars grown in that region?
6. Can the Committee provide advice on which methodological issues and challenges raised in this paper should be given highest priority for future research efforts?

CONTENTS

The role of the Methodology Advisory Committee (MAC) is to review and direct research into the collection, estimation, dissemination and analytical methodologies associated with ABS statistics. Papers presented to the MAC are often in the early stages of development, and therefore do not represent the considered views of the Australian Bureau of Statistics or the members of the Committee. Readers interested in the subsequent development of a research topic are encouraged to contact either the author or the Australian Bureau of Statistics.

METHODOLOGICAL APPROACHES FOR UTILISING SATELLITE IMAGERY TO ESTIMATE OFFICIAL CROP AREA STATISTICS

Jennifer Marley, Daniel Elazar and Kate Traeger
Analytical Services Branch

ABSTRACT

A key strategic goal for the ABS is to exploit emerging sources of data to either partially replace, supplement or validate existing data collections in order to reduce costs, improve quality and produce more responsive and relevant statistical products. Given this imperative, there is a pressing need to assess the available techniques for analysing "big data" problems in terms of their quantifiable statistical reliability and computational feasibility.

In October 2013, methodological research effort commenced at the ABS to investigate the viability of using satellite imagery data to estimate crop area statistics. Other Australian organisations that have decades of experience in analysing satellite imagery data have been pursuing classification of satellite data at the crop type level, but the lack of adequate ground-truth data available in Australia, from which classification methods can learn, has been a major obstacle. The unit record level data that the ABS regularly collects as part of the Rural Environment and Agricultural Statistics Program, however, has the potential to provide a rich source of reference data to train classifiers.

As well as formulating the statistical problem of crop area estimation with satellite imagery data, we outline in this paper the challenges in working with satellite imagery data and present our proposed methodological approaches for estimating crop area statistics, which include four classification methods; namely, support vector machines, Gaussian maximum likelihood classification, classification with kernel density estimation and multinomial logistic regression. We also propose a statistical framework for estimating the bias and variance of crop area estimates that are calculated using crop type predictions for each pixel, given that our ultimate goal is to release these crop area estimates as official statistics. Methods for quantifying the statistical error of such estimates appear to be lacking in the satellite imagery analysis literature, given the focus on prediction.

1. INTRODUCTION

At the beginning of the 2013/14 financial year, the Methodology and Data Management Division (MDMD) in the Australian Bureau of Statistics (ABS) embarked on a new research program consisting of several flagship projects, one of which is the Big Data Flagship Project (BDFP). The goal of the BDFP is to coordinate research and development (R&D) effort that builds a sound methodological foundation for organisational capability in utilising "big data" for our production of statistical outputs (Tam and Clarke, 2014). Tam and Clarke (2014) propose a general Bayesian inferential framework, in terms of conceptualised transformation, sampling and censoring processes applied to the "big data", and argue that, under certain sampling and censoring ignorability conditions, inference can be drawn for the population of interest as if the "big data" were a random sample.

Within the BDFP are a number of targeted R&D initiatives, or work packages, which each seek to meet an identified business need by discovering the value of as yet untapped sources of "big data" and/or advanced analytical techniques. One such work package is concerned with exploring the prospect of supplementing the data collected via traditional means for the ABS Rural Environment and Agricultural Statistics Program with satellite imagery data. We hypothesise that satellite imagery data has the potential to be valuable and useful for:

- validation purposes throughout the agricultural statistics production processes, thereby improving accuracy;
- improving the timeliness and increasing the frequency of agricultural statistical products; and
- creating new and richer agricultural statistical products (for example, it may be possible to forecast future crop yields by developing statistical models that synthesise and analyse satellite imagery data and weather data).

In September 2013, support for this research was obtained from the relevant business area of the ABS, the Rural Environment and Agricultural Statistics Branch (REASB). REASB representatives, with an awareness of the challenges and the time required to develop the necessary capabilities, confirmed that the utilisation of satellite imagery data is part of their long-term vision for the Rural Environment and Agricultural Statistics Program to adopt an improved information model for agricultural statistics, including innovative methodologies, to enable the program to be more responsive to user information demands (Henderson and Pitchford, 2013).

Given the exploratory nature of the work packages under the BDFP, an iterative, or stage-gate, management approach has been adopted to ensure that non-performing initiatives are stopped as early as possible, while promising lines of research receive increased commitment with each successful iteration. As a first step for the satellite imagery data work package, a specific problem was identified to focus initial research effort; that is, to replicate published crop area estimates, to within acceptable statistical error bounds, from the Agricultural Census 2010/11 by accurately classifying satellite imagery data to crop types. In working to solve such a problem, valuable experience in acquiring, handling and analysing satellite imagery data would be gained, as well as an indication of the feasibility of and, therefore, worthiness of continued investment in the methodological research needed for achieving the ultimate goal of producing official agricultural statistics from satellite imagery data.

The purpose of this paper is to formally set out crop area estimation using satellite imagery data and proposed methods for classifying pixels at the crop type level. We highlight the practical and methodological issues we have faced so far in using and analysing satellite imagery data and share ideas and thoughts about future directions for addressing some of these issues. We do not provide solutions to all of the challenges we identify but do seek from MAC advice on which lines of enquiry are best for us to focus our energy and suggestions for other methods that are likely to yield promising results. The implementation of production systems for processing and analysing satellite imagery data in the ABS Rural Environment and Agricultural Statistics Program is beyond the scope of this paper.

The remainder of the paper is organised as follows. Section 2 provides a basic introduction to satellite imagery data; what it is, what it looks like, how it is collected and processed and what the common issues and limitations of satellite imagery data are. Section 3 covers a brief environmental scan of satellite imagery data usage in Australia and other National Statistics Organisations (NSOs). Section 4 formulates the statistical estimation problem we wish to address and Section 5 outlines the four classification methods we have investigated so far. Section 6 explains the process we currently undertake to create the training data we need to estimate the parameters of our classifiers and the test data we need to assess the performance of our trained classifiers. Concluding remarks are given in Section 7, along with our suggestions for our long-term research directions provided the results of our iterative research approach continue to be encouraging.

2. SATELLITE IMAGERY DATA

Remote sensing is the science of obtaining information about an object, area or phenomenon by acquiring data via a measurement or sensor device that does not have direct contact with the object, area or phenomenon under investigation, and then analysing that data (Lillesand, Kiefer and Chipman, 2008). The particular remote sensing of interest for this paper is electromagnetic remote sensing of the Earth's surface and resources via electromagnetic energy sensors mounted on spaceborne platforms, or satellites, that collect data on the levels of reflectance of electromagnetic energy for various features on the Earth's surface.

In what follows, we will briefly describe how the spectral reflectance data is collected and the various registration and correction processes applied to it in order to transform it into an image that truly reflects the scene on the Earth's surface, and then discuss the data's main characteristics and limitations. Throughout this paper, we will refer to the registered and corrected data that we can analyse for our stated purposes as *satellite imagery data*.

2.1 How satellite imagery data is collected and processed

Any energy coming from the Earth's surface, whether it is reflected sunlight or upwelling energy from the Earth itself due to its own finite temperature, can be used to form an image. The general process of electromagnetic remote sensing of the Earth's surface is as follows:

a) energy from the sun is propagated through the Earth's atmosphere;

b) energy interacts with the Earth's surface features;

c) energy is retransmitted through the Earth's atmosphere;

d) sensors mounted on satellites orbiting the Earth measure the levels of emitted and reflected energy;

e) sensor data in pictorial or digital format is generated and transmitted back to ground reception; and

f) sensor data is registered, corrected and transformed into consumable formats and information products that are distributed to end users for analysis and decision making (Lillesand *et al.*, 2008, Richards, 2013).

Sensors on a satellite can measure the level of reflectance of different types of electromagnetic energy, which we can characterise by their wavelength location or wavelength range (which we will call *bandwidth*) on the electromagnetic spectrum (Lillesand *et al.*, 2008). Examples of different types of electromagnetic energy include visible light (in particular, blue, green and red light), near infrared, mid infrared, thermal infrared (that is, heat) and the microwave portion of the spectrum. The atmosphere can have a profound effect on which spectral bandwidths of energy are available to the satellite sensors and the intensity of the energy, due to atmospheric *scattering* and *absorption*; these two phenomena affect points a)–d) in the general remote sensing process described above.

In part b) of our general remote sensing process, it is worth pointing out how electromagnetic energy can interact with features on the Earth's surface. Depending on the material type and condition of the surface feature, the amount of energy reflected, absorbed and transmitted will vary. The amount that surface features reflect, absorb and transmit energy at different bandwidths also varies. These differences and variations are what we rely on to be able to distinguish different features and groundcover types on the Earth's surface (Lillesand *et al.*, 2008).

Once the sensor data is transmitted back to ground reception in part (e), it undergoes a number of corrections and transformations in part (f), a lot of which are mathematically complex, to fix various errors and distortions that arise throughout the remote sensing and measurement process. Sensor data can contain errors that can be categorised under two broad types:

- *Geometric error* – this type of error arises from the relative motions of the satellite, its sensors and scanners, and the Earth causing skewness in the image, and irregularities in the sensors, the curvature of the Earth and uncontrolled variations in the position, altitude, velocity and attitude of the satellite can lead to geometric distortions in the image of varying degrees of severity; and
- *Radiometric error* – this type of error refers to the errors in the measured brightness values of the pixels per spectral bandwidth caused by the effects of the atmosphere as a transmission medium through which the reflected radiation energy must travel, and instrumentation effects, such as calibration differences amongst sensors (Richards, 2013).

Images can also undergo registration with respect to other images taken for the same area of land but at different times so that pixel by pixel comparisons over time can be made.

We will assume that the satellite imagery data created by the end of the general remote sensing process given above, is free of bias and errors when we use it to estimate crop area statistics.

2.2 Characteristics and limitations of satellite imagery data

Satellite imagery data is available as separate images, or scenes, that cover a particular area on Earth. Each scene contains a large number of pixels, which are characterised by the set of spectral reflectance measurements the satellite sensors capture. If there are a sufficiently large number of fine bandwidth measurements captured, then the full spectral reflectance curve or profile of the land to which the pixel corresponds can be reconstructed (Lillesand *et al.*, 2008). An example of typical spectral reflectance profiles for vegetation, water and soil can be viewed in Figure 1 in Jia, Kuo and Crawford (2013).

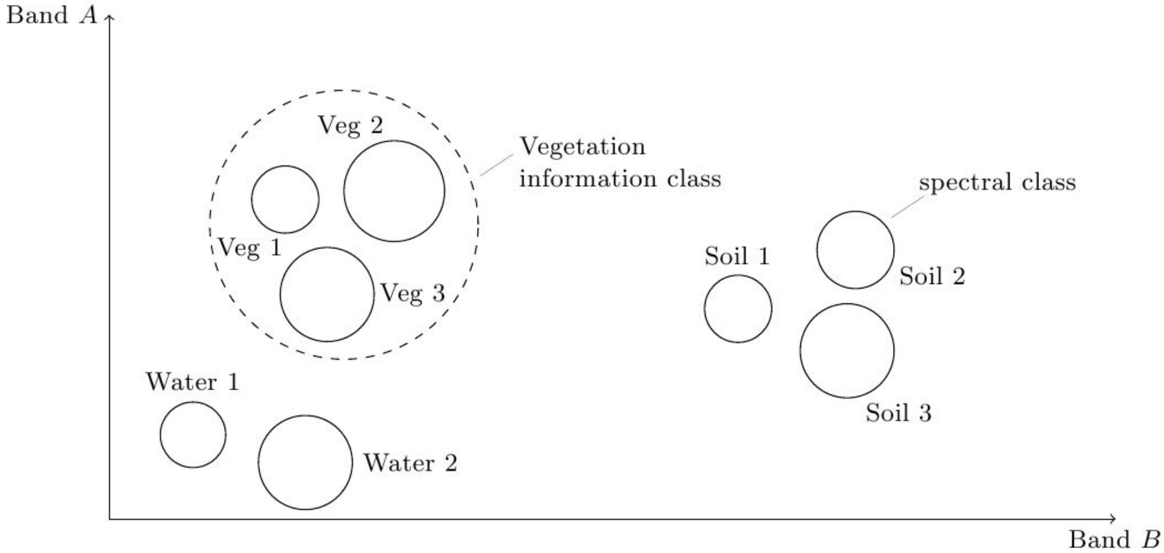
Given the spectral reflectance measurements associated with each pixel, the process of classification is essentially a mapping of those measurements to a label that identifies a particular ground-cover type for the land the pixel represents.

If the number of spectral reflectance measurements captured by the sensors on a satellite for each pixel is in the order of hundreds, called *hyperspectral data*, then it seems feasible that the ground cover type could be predicted with high confidence. *Multispectral data*, where each pixel has spectral reflectance measurements in the order of tens, makes confident classification less feasible, as distinguishing features of the full spectral reflectance profile may not be captured.

The satellite imagery data available to us that covers the same period of time as the Agricultural Census 2010/11 is that of Landsat-7, as disseminated by the U.S. Geological Survey (USGS) (USGS, 2014). Landsat-7 was launched on April 15, 1999 into a repetitive, circular, sun-synchronous, near-polar orbit (Lillesand *et al.*, 2008) and is still active. The sensor on board is the Enhanced Thematic Mapper Plus (ETM+) which collects measurements for the bandwidths detailed in National Aeronautics and Space Administration (2014). Reflectance measurements are collected at a resolution of 30m for six bandwidths in the visible, near infrared and mid infrared spectral regions and at a 60m resolution for the seventh bandwidth in the thermal range. The orbit results in a 16-day repeat cycle, meaning that satellite images are captured for the same area of land every 16 days (USGS, 2013).

Since Landsat-7 captures seven spectral reflectance measurements, we are thus currently limited to multispectral data. A vector of the seven reflectance measurements allows us to work with a set of data points in seven dimensional space. A coordinate system with as many dimensions as there are reflectance measurements in the pixel vector is often called *spectral space* in the remote sensing literature (Richards, 2013). Data points in the spectral space corresponding to pixels that represent land with different ground cover types cluster in different regions of the spectral space. These clusters are called *spectral classes*. Separate spectral classes can group in neighbourhoods, which are called *information classes*. Figure 2.1 presents a diagrammatic representation of spectral and information classes in a two dimensional spectral space, and they are shown as distinct clusters and groups of clusters, respectively. In reality, especially for our research problem where the spectral classes we wish to distinguish relate to different crop types, it is more likely that clusters overlap and are part of a continuum of data in spectral space (Richards, 2013), causing the classification problem to be a difficult one.

2.1 Spectral and information classes in two dimensional spectral space



Spectral and information classes in two dimensional spectral space defined by two spectral bandwidths, A and B. Circles with solid lines depict clusters of data points in spectral space that belong to the same spectral class while the circle with a dotted line represents a broader neighbourhood or cluster of spectral classes, called an information class.

Methods for handling the classification problem fall under two broad types – supervised classification and unsupervised classification. Supervised classification methods use a sample set of pixels that we have already labelled, using another source of information, as *training data* from which to learn. Training of these methods or classifiers refers to the estimation of the parameters that the classifier needs in order to be able to recognise and predict the most likely label for other unknown pixels. Unsupervised classification is an umbrella term for methods that assign pixels in an image to spectral classes without foreknowledge about the existence or labels of those classes (Richards, 2013). It is often performed using clustering type methods.

There is a third approach for handling the classification problem, a hybrid approach which is a combination of unsupervised and supervised classification techniques being used together. For example, unsupervised classification can be used to determine a spectral class decomposition of the image data to inform the subsequent application of supervised classification techniques (Richards, 2013).

Our research efforts so far have concentrated on four supervised classification methods, which are detailed in Section 5. In order for us to test and evaluate the classification methods, we have needed to create a set of training data; our current process for and challenges with creating quality training data are discussed in Section 6.

The supervised classification methods we have looked at use the spectral reflectance measurements of the image pixels only but given the nature of the data and the research question we have in mind, it is clear that the performance of the classification methods would benefit with the inclusion of spatial and/or time series analysis techniques. This idea is not directly considered in this paper but discussed as part of our future research directions in Section 7.

Some notable issues with and limitations of satellite imagery data, that present challenges for its utilisation in the estimation of crop area statistics, include:

- *Missing and contaminated data due to cloud cover* – obvious clouds in an image can render pixels of interest to be considered as missing values while thin cloud cover, which is hard to detect, can lead to contaminated spectral reflectance measurements for the affected pixels;
- *Missing or poor quality data due to on board satellite equipment failures* – for example, a failure in the scan line corrector on the Landsat-7 satellite back in May 2003 caused significant levels of missingness (approximately 22%) in Landsat-7 images ever since (USGS, 2013). The scan line corrector compensates for the forward motion of the spacecraft so that the resulting scans are aligned parallel to each other. Without it, the sensors scan the Earth in a "zig-zag" fashion, leading to reflectance measurements not being detected for some areas and some areas being imaged twice. This results in images with black data gaps that look like alternating wedges and that increase in width towards the edge of the image. Although this is not a problem for the next satellite in the Landsat series, Landsat-8, it highlights the risk of future failures in the on board satellite equipment that could affect the quality of the data it collects; and
- *Insufficient image resolution* - an inherent assumption in the classification methods we have considered so far is that each pixel contains only one type of ground cover, or crop, that we wish to predict. Given that the area of land each pixel represents in a Landsat-7 image is 30m × 30m, this assumption is not valid when the variability of the ground cover is high (Jones and Vaughan, 2010). Appendix A contains images that demonstrate the pixels in Landsat-7 images can contain a mix of ground cover types.

We have not directly addressed these issues in this paper but offer thoughts and ideas about future research directions for overcoming them and other challenges in Section 7.

3. ENVIRONMENTAL SCAN

3.1 International efforts in utilising satellite imagery data for official agricultural statistics

A number of NSOs around the world are embarking on similar research paths to the ABS in terms of using satellite imagery data to aid in the production of official agricultural statistics. We briefly describe the efforts of four other NSOs below, all of which have varying research questions, methodological approaches, experience with managing and analysing satellite imagery data and levels of progress and success.

3.1.1 National Agricultural Statistics Service, U.S. Department of Agriculture

The National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture (USDA) is probably considered the world leader in using satellite imagery data to produce crop area statistics.

They produce the Cropland Data Layer (CDL) product, which is a raster-formatted, geo-referenced, crop-specific land cover map based on medium resolution satellite imagery, Farm Service Agency (FSA) Common Land Unit (CLU) data, which contains extensive agricultural ground truth data, and other ancillary data, such as the National Land Cover Dataset for non-agricultural ground truth data. A decision tree supervised classification method, via Rulequest Research's commercial software See5 Decision Tree, is employed to derive state-level decision trees and predict state-level crop cover classifications. The satellite images used as training data for the state-level decision tree classifiers are selected based on optimal dates for separation of crop types (Boryan, Yang, Mueller and Craig, 2011).

To calculate crop acreage estimates, NASS have found that intuitive pixel counting estimates consistently underestimate the actual acreage (that is, NASS official estimates) and so they model the relationship between official survey weighted estimates and the pixel counting estimates using simple linear regression to improve upon the satellite imagery pixel based estimate (Boryan *et al.*, 2011).

Accuracy of the CDL product is based on Kappa coefficients, which measure the difference between the agreement in the accuracy matrix and the agreement that could be expected to occur by chance (Congalton and Green, 1999). Accuracies are generally 85–95% correct pixels for major crop categories. Although not generally available, NASS also produce a classification confidence layer for CDL. The confidence values it contains don't reflect the accuracy of a given pixel's classification but rather a measure of how well the decision to identify the pixel within a specific category fit within the decision tree rule set. Further details in Liu, Gopal and Woodcock (2004) reveal that the confidence measure does not have a statistical inferential basis.

In 2011, NASS released CropScape, a highly accomplished interactive web CDL exploring system that allows users to query, visualise, download and analyse CDL data geospatially in a freely accessible online environment (Han, Yang, Di and Mueller, 2012).

3.1.2 Statistics Canada

Statistics Canada began a research program in 2013 to investigate the feasibility of using satellite imagery data to aid in modelling crop yield and estimating crop area statistics (Bedard and Reichert, 2013).

They have tested, with varying degrees of success, linear regression models to predict crop yields where the covariates included in the model were crop yield estimates based on survey data from previous years, climate information and the Normalised Difference Vegetation Index (NDVI) calculated from satellite imagery data. The NDVI is used to detect the presence and condition of vegetation and is the ratio of the difference and sum of near-infrared and visible red measurements (Jones and Vaughan, 2010).

For crop area estimation, they have performed similar experiments to the ABS in that they are comparing estimates based on predictions from classifiers applied to satellite imagery data to estimates resulting from an area sample frame methodology. Currently, work is being trialled in the southern part of the province of Manitoba.

One of their greatest learnings so far is that collecting sufficient ground truth data is paramount for the success of quality agricultural estimates being produced from satellite imagery data.

3.1.3 Centraal Bureau voor de Statistiek (Statistics Netherlands)

Centraal Bureau voor de Statistiek (CBS) carried out a pilot project in late 2012 and early 2013 to investigate how satellite imagery data could be used to calculate arable crop harvests. They collaborated on this pilot project with a private organisation, eLEAF, who have specialist knowledge in interpreting satellite imagery data.

Two spatial datasets, containing the volume of cereal and potato harvests per mapping square over the whole of the Netherlands for 2012, were created by applying eLEAF's well-established and validated model for calculating dry matter biomass production that uses satellite imagery and meteorological data (Meurink, 2013). By overlaying this harvest data map with a complete farmland parcel register for 2012, it was possible to calculate harvested area and volume for each crop per farm and compare these to the corresponding values reported by respondents to the CBS's Arable Harvest Projection Survey, 2012. As we intend to do with crop areas reported by respondents to the ABS's Agricultural Census, CBS assume the harvest records submitted by respondents to their survey reflect reality. Differences between the harvest per hectare recorded by the survey respondents and the calculations based on satellite imagery data were too large for CBS to be confident that they could calculate reliable official harvest statistics at the national and province levels using satellite imagery data (Meurink, 2013).

As this was an initial pilot project, many shortcuts were taken that potentially induced needless inaccuracies. CBS are currently exploring with eLEAF promising options for improving upon this research.

3.1.4 National Bureau of Statistics, China

In 2003, in order to improve timeliness and accuracy of agricultural statistics needed to inform food policy and national economic planning, the National Bureau of Statistics (NBS), China initiated discussions with relevant academic and research institutes about the possibility of combining the national agricultural statistical survey system and remote sensing technology. This led to research and tests that resulted in summer grain crop area estimates for a number of provinces in China using remote sensing technology and traditional statistical methods. In 2006, the Ministry of Science and Technology established the first key project of the *National High Technology Research and Development Program 863* in the field of "Earth Observation and Navigation Technology: Research and Application of Key Technologies of National Statistics and Remote Sensing Service System" (Zhang, Zhu, Pan, Hu and Zhang, 2010). This led to the development of the National Statistics and Remote Sensing System of Crop Production (NSRCP). The goal of the NSRCP was to produce very accurate estimates of crop area and yield statistics for the major crops, wheat, corn and rice, at provincial level and county level by integrating a number of data sources including satellite imagery data, bio-meteorological data, traditional sample survey data and statistics from previous years. While it addresses a number of the problems a system such as this is likely to have, in order for the NSRCP to produce estimates of sufficient accuracy and timeliness, Zhang *et al.* (2010) suggest that a wireless sensor network needs to be established to measure crop conditions in real time and crop area and yield estimation methods using multi-source, multi-scale satellite imagery data and real-time crop monitoring data from the proposed sensor network need to be improved.

3.2 Australian efforts in utilising satellite imagery data

There are a number of agencies and organisations around Australia that work with, analyse and make use of satellite imagery data for a number of different purposes.

Geoscience Australia (GA) is Australia's principal earth resource satellite ground station and data processing facility. They have decades of experience in managing, processing, analysing and distributing satellite imagery data of Australia and work in collaboration with international satellite operators. GA make use of satellite imagery data for a range of applications and are involved in a number of projects to provide information to decision makers on topics related to the environment, agriculture and community safety.

Most relevant to current ABS satellite imagery research effort is GA's Dynamic Land Cover Dataset (DLCD), which uses satellite imagery data from the NASA Moderate Resolution Imaging Spectroradiometer (MODIS) satellite. It is designed to be a nationally consistent thematically comprehensive land cover reference for Australia; the first of its kind. It provides land cover information at 250m × 250m resolution over the period April 2000 to April 2008, using 34 land cover categories. Different vegetation land cover types were distinguished using the Enhanced Vegetation Index and then an innovative time series analysis technique was applied at the pixel level which reduced each pixel's time series to 12 coefficients based on statistical, phenological and seasonal characteristics. A support vector clustering algorithm was then used to cluster the coefficients and the resultant classes were labelled using catchment scale land use mapping data (which is discussed below) and the National Vegetation Information System (Lymburner, Tan, Mueller, Thackway, Lewis, Thankappan, Randall, Islam and Senarath, 2011).

Since early 2013, GA have also been working on the creation of a 'cube' of Earth observation datasets by stacking Landsat image 'tiles' in time sequences covering the same area of ground (National Computational Infrastructure, 2013). By standardising and registering fifteen years of Landsat-5 and -7 imagery that covers the entire continent of Australia, GA have created the Data Cube, which is available to be analysed via the National Computational Infrastructure and allows for more sophisticated quantitative analysis. We plan to use the Data Cube in the future.

The Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES) produce two types of land use data, (1) catchment scale land use mapping and (2) national scale land use mapping data. Catchment scale land use mapping is based on the integration of land tenure and other types of land use data, fine-scale satellite imagery data and information collected in the field. National scale land use mapping is based on coarse-scale satellite imagery data (that is, pixels of size 1.1 square kilometres), ABS agricultural statistics and ground control point data for agricultural land uses, and various other digital maps, including the finer resolution catchment scale land use data, for non-agricultural land uses (ABARES, 2011).

ABARES also manage the Ground Cover Monitoring for Australia project, which is a collaborative partnership between the Department of Agriculture, Fisheries and Forestry, the Commonwealth Science and Industrial Research Organisation (CSIRO), the Terrestrial Ecosystem Research Network and the state and territory governments. The project delivers estimates on ground cover at 500m resolution from MODIS satellite imagery data using the method in Guerschman, Hill, Renzullo, Barrett, Marks and Botha (2009). A national network of field sites with group cover measurements was also established to validate and improve upon this satellite imagery analysis method (Stewart, Rickards and Randall, 2013).

Different parts within CSIRO use satellite imagery data for a large range of applications; we will mention some of the projects and initiatives they are involved in that are most relevant to our work. The Environmental Earth Program is using remote sensing to investigate changes in land surface-climate interactions, coupling of water and carbon balances and the observation and prediction of hydrological processes (CSIRO, 2011). CSIRO Marine and Atmosphere Research (CMAR) coordinate the AusCover facility, which is a national expert network that provides remote sensing data time series and satellite imagery based biophysical map products for Australia. CMAR are also responsible for the Atmosphere and Land Observation and Assessment Program which processes and analyses satellite imagery data for the measurement of land surface properties, such as soil moisture and vegetation. This data is assimilated into weather and climate forecast models and feeds into forecasting systems used by the Bureau of Meteorology (BoM) (CSIRO, 2011). The Centre for Australian Weather and Climate Research (CAWCR) is a partnership between CSIRO and BoM of which the common goal is to develop Australia's next generation climate model, ACCESS. ACCESS is tested with and fed satellite imagery data and data from other climate sensors.

A number of other groups and organisations in Australia are active in satellite imagery data research and application, including:

Bureau of Meteorology (BoM)

> The Bureau of Meteorology are extensive users of data from multiple satellite sources. Notable examples of how they utilise satellite imagery data include analysing sea surface temperature for ocean forecasting and prediction of tropical cyclones and other severe weather events, and calculating and mapping the NDVI annually to quantify and visualise the vegetation state of Australia relative to the long-term average.

University of New South Wales (UNSW), Canberra Campus

> The School of Engineering and Information Technology at UNSW, Canberra, has a remote sensing research group who focus their research efforts on improving methods for handling the three main issues in remote sensing data interpretation – data compression and transmission, data correction and data analysis.

Landgate – Satellite Remote Sensing Services

Landgate is the primary source of land information and geographic data in Western Australia. For over a decade, Landgate has used a combination of NASA and in-house software to detect bushfire hotspots from MODIS satellite imagery data. These hotspots are used as ignition points for University of Western Australia developed software, *Australis*, that simulates bushfire behaviour and spread over various fuel types found in Australia. Landgate also provide a range of online farm-related data products based on satellite imagery data that show the variability between cropped and pasture paddocks, track vegetation loss and measure crop growth rates.

Curtin University – Remote Sensing and Satellite Research Group (RSSRG)

The five main research streams of RSSRG are Hyperspectra, Atmosphere, Marine and Estuarine, Terrestrial and Underwater imaging. Their research work quantitatively analyses well-calibrated spectral radiometric observations to generate products with validated accuracy.

Terrestrial Ecosystem Research Network (TERN)

TERN, at the University of Queensland, connects ecosystem scientists and enables them to collect, contribute, store, share and integrate data across disciplines. TERN operates as a network of facilities in partnership with other agencies. Such facilities include AusCover with the CSIRO, the Eco-Informatics facility, which combines spatial, satellite, ecological and genomics data (TERN, 2012), and the Soil and Landscape Grid of Australia facility, which uses a combination of remote soil sensing data and other soils data to estimate key soil attributes at a scale relevant to ecosystem process. TERN have also partnered with CSIRO and Google to put detailed satellite imagery of Australian landscape into Google Earth (TERN, 2012).

4. STATISTICAL FORMULATION OF CROP AREA ESTIMATION USING SATELLITE IMAGERY DATA

Given that the ultimate goal of our current research effort in using satellite imagery data is to produce crop area estimates of sufficient quality to be published as official ABS statistics, we propose the following formulation of the statistical problem we wish to address.

Let U_t be the complete set of satellite imagery pixels that cover the areas of Australia used for agricultural purposes at time t, and $|U_t| = N_t$. Pixels that are associated with metropolitan cities, urban areas, townships, deserts, waterways, national parks and natural bush areas, to name a few, are out of scope. Our formulation of the crop area values we wish to estimate and our proposed estimator for these values, along with the classification methods we describe in Section 5, will appropriately account for this situation. This area based scope may be identified by data sources such as ABARES national scale land use maps as mentioned in Subsection 3.2, the production of which involves the development of a non-agricultural land use mask (ABARES, 2011). The utility of these data sources will need to be assessed in terms of availability, timeliness and consistency with our datasets and statistics.

Let $\Psi_t = \{\psi_g : g = 1,\ldots,G_t\}$ be the set of geographic areas, indexed by g, at time t for the desired level of statistical output, and $|\Psi_t| = G_t$. For example, if we wish to produce estimates of land area used to grow wheat at the state level, then the elements of Ψ_t will be the eight states and territories of Australia. We have, again, made the set of output geographic areas time-specific as the definitions and boundaries of these areas may change over time. For example, National Resource Management regions (NRMs) are administrative regions that are commonly used for environmental and agricultural reporting and for which the boundaries are occasionally revised (ABS, 2011).

Let $\Omega_t = \{\omega_c : c = 1,\ldots,C_t\}$ be the set of crop type labels, indexed by c, that are of interest for publication and known to be grown in Australia at time t, and $|\Omega_t| = C_t$. We assume that it is necessary to make the set of all crop types grown in Australia to be time specific as crops are grown depending on seasonal, climatic and weather conditions and possibly also economic conditions and activities. We will also assume that there are auxiliary sources of information that can be used to identify the crop types that constitute this time-specific set.

If in the future we decide to employ region-specific classifiers for practicality and efficiency reasons, then it may be more appropriate to define $\Omega_{\upsilon_r t} \equiv \Omega_{rt} \subseteq \Omega_t$, the set of all crop types known to be grown in region υ_r at time t, and $|\Omega_{rt}| = C_{\upsilon_r t} \equiv C_{rt} \le C_t$, for all $\upsilon_r \in \Upsilon_t$, where $\Upsilon_t = \{\upsilon_r : r = 1,\ldots,R_t\}$ is the set of all regions, indexed by r, covering the Australian continent at time t, and $|\Upsilon_t| = R_t$.

Deciding what these regions should be to optimise the performance of the region-specific classifiers is a possible line of future research.

For ease of notation, we will drop the index t, and c, g and r will be used in place of ω_c, ψ_g and υ_r, respectively, in subscripts for the remainder of the paper.

We can express the total area of land used to grow crop ω_c in geographical output area ψ_g as

$$T_{\omega_c \psi_g} \equiv T_{cg} = \sum_{i=1}^{N_g} a_i Z_{ic} , \qquad (4.1)$$

where:

- $U_g \subseteq U$ is the complete set of pixels that cover land used for agricultural purposes in geographical output area ψ_g, and $\left|U_g\right| = N_g$;
- a_i is the area of land that pixel i represents (for example, $a_i = 900m^2$ for all i in Landsat 7 images (USGS, 2013)); and
- Z_{ic} is an indicator variable for pixel i, which takes the value 1 if the crop type growing on the land to which pixel i corresponds is ω_c, otherwise, it takes on the value 0.

We propose the following estimator for T_{cg},

$$\hat{T}_{cg} = \sum_{i=1}^{N_g} a_i \hat{Z}_{ic} , \qquad (4.2)$$

where N_g and a_i are defined as before and $\hat{Z}_{ic}$ is an estimator for Z_{ic}. $\hat{Z}_{ic}$ is generated according to some classification method that labels pixels as different crop classes depending on their observed reflectance values. The label predictions this classification method generates are subject to a certain underlying probability distribution. We can express this algebraically as,

$$\hat{Z}_{ic} = I\left(\hat{y}_i = c\right) \text{ and } \hat{y}_i = \hat{\mathcal{F}}\left(\underset{\sim}{x}_i ; \hat{\Theta}\right), \qquad (4.3)$$

where:

- $I(\cdot)$ is an indicator function such that if the argument is true, it takes the value 1, otherwise it takes the value 0;
- $\hat{y}_i \in \{1, \ldots, C\}$ is a predicted crop class for pixel i, which has a true crop class $y_i \in \{1, \ldots, C\}$;
- $\hat{\mathcal{F}}\left(\cdot\,; \hat{\Theta}\right)$ is an estimate of a true assignment mechanism $\mathcal{F}(\cdot\,; \Theta)$ that is controlled by a set of parameters, Θ; and

- $\underset{\sim}{x}_i$ is an m dimensional vector containing the m reflectance measurements captured by satellite sensors for pixel i (for example, $m = 7$ for Landsat-7 satellite imagery data (USGS, 2013)).

If we ignore the potential biases and errors that could be induced in the satellite images we analyse due to the collection and correction processes mentioned in Subsection 2.1, then the only source of bias or error in (4.2) is $\hat{Z}_{ic}$.

The bias of $\hat{T}_{cg}$ is given by,

$$\mathcal{B}\left(\hat{T}_{cg}\right) = \sum_{i=1}^{N_g} a_i \mathcal{B}\left(\hat{Z}_{ic}\right), \tag{4.4}$$

and the variance of $\hat{T}_{cg}$ is given by,

$$\operatorname{Var}\left(\hat{T}_{cg}\right) = \sum_{i=1}^{N_g} a_i^2 \operatorname{Var}\left(\hat{Z}_{ic}\right) + \sum_{i=1}^{N_g} \sum_{j \neq i}^{N_g} a_i a_j \operatorname{Cov}\left(\hat{Z}_{ic}, \hat{Z}_{jc}\right). \tag{4.5}$$

The expression for the variance of $\hat{T}_{cg}$ can be simplified further if the classification method used classifies pixels independently of other pixels (that is, solely on the basis of the $\underset{\sim}{x}_i$'s), which would mean $\operatorname{Cov}\left(\hat{Z}_{ic}, \hat{Z}_{jc}\right) = 0$ for all $i \neq j$.

As described in Subsection 2.2, supervised or unsupervised classification methods, or alternatively, some kind of supervised-unsupervised hybrid approach can be used to label a pixel with a crop type and thus provide an estimate for $\hat{Z}_{ic}$. We have only considered supervised classifiers so far in our research. For supervised classifiers to 'learn', that is, to estimate the parameters, Θ, of the classifier $\mathcal{F}\left(\cdot\,;\Theta\right)$ in order to use it to label unknown pixels, we need a set of training pixels, or training data points, $\mathcal{T} = \left\{(y_i, \underset{\sim}{x}_i) : i = 1, \ldots n\right\}$, where $\left|\mathcal{T}\right| = n$.[1]

In Section 5, we will briefly describe the four supervised classification methods we have considered so far to estimate $\hat{Z}_{ic}$ and discuss the bias and variance properties of $\hat{Z}_{ic}$ and, consequently, our crop area estimator $\hat{T}_{cg}$, under each of these methods.

[1] If we wish to develop region-specific classifiers, then they will each need their own set of training pixels $\mathcal{T}_{rt}$ and $\left|\mathcal{T}_{rt}\right| = n_{rt}$. It is not a necessary condition that $\mathcal{T}_{rt} \subseteq U_{rt}$, the set of all pixels in region υ_r, nor, consequently, that $n_{rt} \leq N_{rt}$, where $\left|U_{rt}\right| = N_{rt}$, as it is likely that in order to represent the variability of the reflectance vectors associated with the different crop types in Ω_{rt} in the training set $\mathcal{T}_{rt}$, training pixels will need to be sourced from other regions and other time periods.

5. CLASSIFICATION METHODS

The four classification methods we have considered so far in our research efforts are support vector machines, maximum likelihood classification, classifiers based on kernel density estimation and multinomial logistic regression. Each will be outlined in turn over the next four subsections, along with the implications of using each of the methods on the bias and variance of our estimator of crop area, $\hat{T}_{cg}$, as given in (4.4) and (4.5), respectively.

5.1 Support Vector Machines

Support Vector Machines (SVMs) are a supervised non-parametric method with a geometric basis that has proven to be a powerful technique for non-linear classification, regression and outlier detection in many application fields. SVMs have their foundation in machine-learning and were first introduced in Boser, Guyon and Vapnik (1992) and were developed for binary classification in Cortes and Vapnik (1995). SVMs have a natural application to the satellite imagery classification problem, as demonstrated in the formative paper Gualitieri and Cromp (1998) where SVMs achieved high prediction accuracy on a difficult satellite imagery classification problem by utilising the full dimensionality of hyperspectral satellite data (that is, hundreds of reflectance measurements per ground pixel). Although SVMs haven't been widely accepted in the statistical community, they have proven to handle complex classification problems efficiently.

5.1 Linear, soft-margin and non-linear support vector machines

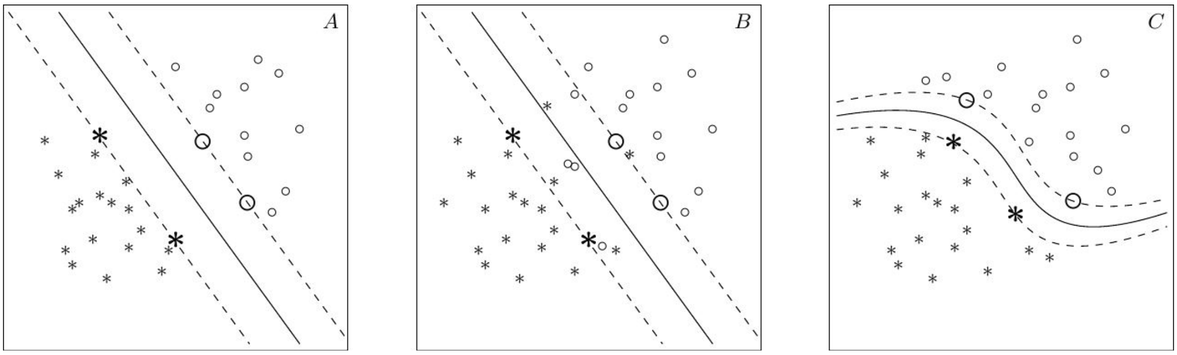


Figure 5.1: Support vector machines in two dimensions where open circles and asterisks represent data points for two different classes. Diagram A demonstrates a linearly separable situation. Diagram B demonstrates the use of a soft margin when the classes overlap. Diagram C demonstrates the use of a nonlinear decision boundary when the classes are nonlinearly separable.

Assuming our satellite imagery data context, the goal of an SVM for binary classification is to define an optimal separating or decision surface in multispectral space that maximises the margin between the two classes' closest training data points. The training data points lying on the margin boundaries are called the *support vectors* (since training data points have an associated vector that describes their location in the spectral space) and the optimal decision surface goes through the centre of the margin. In figure 5.1, a solid line represents an optimal decision surface, dotted lines represent margin boundaries and the larger, bolded data points are the support vectors.

Figure 5.1 (A) shows a linearly separable case in two dimensions. If the training data points of the two classes overlap, as shown in figure 5.1 (B), then a *soft-margin* SVM (Cortes *et al.*, 1995) can be used in which the training data points that fall on the 'wrong' side of the margin are weighted down in the optimisation procedure to reduce their influence.

Linear SVMs are algorithms that depend on the input training data only through dot products, making them a part of a more general category of kernel methods. SVMs also have great capability to handle non-linearly separable classes. By replacing every dot product with a nonlinear kernel transformation function, the SVM algorithm can fit the maximum margin hyperplane, that is, a linear decision surface, in the transformed higher dimensional feature space which actually generates a non-linear decision boundary in the input space, as shown in figure 5.1 (C). Kernel transformation functions tend to spread out the data points in this higher dimensional feature space which facilitates a linear separating surface being able to be found between the two classes. This technique is called the *kernel trick* and it allows us to never explicitly work in the higher dimensional feature space (Gualitieri *et al.*, 1998); we are never confronted with the cost of computing the large number of vector components in that space but it is a way to find a non-linear decision surface using a method designed for a linear classifier. SVMs, and kernel methods in general, have seen rapid development over the last two decades, particularly in remote sensing applications (Richards, 2013).

The following provides an overview of the mathematical formulation of a binary SVM classifier in our satellite imagery data context, assuming the simple case of linearly separable classes. Details of the extensions to the linear SVM to allow for overlapping classes and non-linearly separable classes are provided in Appendix B.

Suppose that we have a set of training data points, $\mathcal{T}$, as defined in Section 4 but where $y_i \in \{1,2\}$ since we are only considering binary classification. To align with the standard formulation in the literature, we will define the SVM in terms of the $\underset{\sim}{x}_i$ and u_i, where $u_i = +1$ if $y_i = 1$ and $u_i = -1$ if $y_i = 2$.

Since we are assuming that the training data points for the two classes are linearly separable in our multispectral or m dimensional space, then our goal is to find a hyperplane, which has a general equation of the form,

$$\underset{\sim}{w}^{\mathrm{T}}\underset{\sim}{x}+b=0\ ,$$

where:

- $\underset{\sim}{x}$ is a point on the hyperplane;
- $\underset{\sim}{w}$ is an m dimensional vector perpendicular to the hyperplane; and
- b is the distance of the closest point on the hyperplane to the origin,

that separates the two classes of training data points and is the maximum distance from the closest training data points of each class; as represented by the solid line in figure 5.1 (A). The corresponding classifier function, or discriminant function, for an unknown pixel l with reflectance vector x_l is,

$$f\left(\underset{\sim}{x}_l;\underset{\sim}{w},b\right)=\operatorname{sgn}\left(\underset{\sim}{w}^{\mathrm{T}}\underset{\sim}{x}_l+b\right). \qquad (5.1)$$

That is, if a new, unseen pixel has a reflectance vector that places it on the same side of the hyperplane as the '$+1$' or positive examples in the training dataset, it will be classified as '$+1$' (that is, $\hat{u}_l=+1\Rightarrow\hat{y}_l=1$) or, conversely, it will be classified as '–1' (that is, $\hat{u}_l=-1\Rightarrow\hat{y}_l=2$) if its reflectance vector places it on the other side of the hyperplane.

In order to find the optimal separating hyperplane, we need to first define the concept of a margin. For a given hyperplane, let x_+ and x_- denote the training data points closest to the hyperplane among the positive and negative classes, respectively. Via a simple geometric argument, the margin of a hyperplane given our set of training data points can be seen to be,

$$\frac{1}{2}\underset{\sim}{\hat{w}}^{\mathrm{T}}\left(\underset{\sim}{x}_+-\underset{\sim}{x}_-\right),$$

where $\underset{\sim}{\hat{w}}=\underset{\sim}{w}/\left\|\underset{\sim}{w}\right\|$ is a unit vector in the direction of $\underset{\sim}{w}$, and we assume that x_+ and x_- are equidistant from the decision boundary, that is,

$$\underset{\sim}{w}^{\mathrm{T}}\underset{\sim}{x}_+ +b=a \text{ and } \underset{\sim}{w}^{\mathrm{T}}\underset{\sim}{x}_- +b=-a\ , \qquad (5.2)$$

for some constant $a>0$ (Ben-Hur and Weston, 2009).

We can scale the parameters of the hyperplane, $\underset{\sim}{w}$ and b, by a constant without changing the hyperplane so that $a = 1$, and by taking the difference between the equations in (5.2) and dividing by $\|\underset{\sim}{w}\|$, we find that the margin is,

$$\frac{1}{2}\hat{\underset{\sim}{w}}^{\mathrm{T}}\left(\underset{\sim}{x}_{+} - \underset{\sim}{x}_{-}\right) = \frac{1}{\|\underset{\sim}{w}\|}.$$

As stated earlier, the goal of an SVM is to find the hyperplane that separates the two classes with the largest margin. We want to maximise the geometric margin $1/\|\underset{\sim}{w}\|$, which is equivalent to minimising $\|\underset{\sim}{w}\|^2$ and leads us to the quadratic optimisation problem given by,

$$\begin{aligned} &\underset{\underset{\sim}{w},b}{\text{minimise}} \quad \frac{1}{2}\|\underset{\sim}{w}\|^2 \\ &\text{subject to:} \quad u_i(\underset{\sim}{w}^{\mathrm{T}}\underset{\sim}{x}_i + b) \geq 1, \quad i = 1,\ldots,n, \end{aligned} \tag{5.3}$$

where the ½ factor in the objective function is cosmetic. The constraint in (5.3) ensures, by multiplying the argument of (5.1) by u_i, that all positive data points are on the positive side of the hyperplane and all negative data points are on the negative side, and that all data points in the training dataset that aren’t the support vectors are at least as far away from the hyperplane as the support vectors.

To solve the optimisation problem in (5.3), the method of Lagrange multipliers can be employed to obtain the *dual representation*, which can be solved numerically. The dual representation is,

$$\begin{aligned} &\underset{\underset{\sim}{\alpha}}{\text{maximise}} \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j u_i u_j \underset{\sim}{x}_i^{\mathrm{T}}\underset{\sim}{x}_j \\ &\text{subject to:} \begin{cases} \text{i.} \quad \sum_{i=1}^{n}\alpha_i u_i = 0, \\ \text{ii.} \quad \alpha_i \geq 0, \\ \text{iii.} \quad \alpha_i\left\{u_i\left(\underset{\sim}{w}^{\mathrm{T}}\underset{\sim}{x}_i + b\right) - 1\right\} = 0, \quad i = 1,\ldots,n, \end{cases} \end{aligned} \tag{5.4}$$

where the α_i are the Lagrange multipliers.

Constraint (iii.) in (5.4) is of interest as it implies, for every training data point i, either $\alpha_i = 0$ or $u_i\left(\underset{\sim}{w}^{\mathrm{T}}\underset{\sim}{x}_i + b\right) = 1$. The latter condition is only true for the support vectors, which means $\alpha_i = 0$ for all training data points that aren’t the support vectors. It seems, then, that most of the training data points are irrelevant, which is true to some degree but must be qualified with the following: there is no way of knowing which of the training data points will be the support vectors until the optimisation problem in (5.4) has been solved.

Once we have found the optimal decision hyperplane and thus the support vectors, we can disregard the rest of the training data points when we wish to classify further, unknown pixels (Richards, 2013). This means that our classifier function for an unknown pixel l in (5.1) becomes,

$$f(\underset{\sim}{x}_l;\alpha_s : s \in \mathcal{S}, b) = \operatorname{sgn}\left(\sum\nolimits_{s\in\mathcal{S}} \alpha_s y_s \underset{\sim}{x}_s^{\mathrm{T}} \underset{\sim}{x}_l + b\right), \tag{5.5}$$

where $\mathcal{S} \subseteq \mathcal{T}$ is the set of support vectors.

This translates to our classifier function for $\hat{y}_l$, $\hat{\mathcal{F}}$, as defined in (4.3), being, under the SVM classification method,

$$\mathcal{F}\left(\underset{\sim}{x}_l;\alpha_{s:s\in\mathcal{S}}, b\right) = \begin{cases} 1 & \text{if } f\left(\underset{\sim}{x}_l;\alpha_{s:s\in\mathcal{S}}, b\right) = +1 \\ 2 & \text{if } f\left(\underset{\sim}{x}_l;\alpha_{s:s\in\mathcal{S}}, b\right) = -1 \end{cases} \tag{5.6}$$

It is concerning that within class distribution is ignored when classifying unknown pixels. For example, if crop classes of interest have underlying long-tailed probability distributions, then this will lead to high variability in location of the support vectors and hence the decision boundary. The use of soft-margin and non-linear SVMs (see Appendix B), along with cleverly designed ensembles of SVM classifiers, may overcome this concern; further research is required to verify these speculations.

The SVM representation in (5.4) only accounts for binary classification while we actually require a classification method that can handle $C \geq 2$ classes. While a brief discussion of some of the methods in the literature for reformulating the SVM quadratic optimisation problem for multiple classes is given in Appendix C, a simple way to handle the multi-class problem is to utilise a number of binary SVM classifiers in conjunction with a voting scheme such as:

- *One-Against-All*: C binary SVM classifiers are trained to separate one class from the rest. An unknown pixel l is classified by applying all binary classifiers to its reflectance vector $\underset{\sim}{x}_l$ and then choosing the label ω_c, that is, setting $\hat{y}_l = c$, where the argument of the sgn function in the classifier function (5.1) for the c-th classifier is largest; or
- *One-Against-One*: $\binom{C}{2} = C\left(C-1\right)/2$ binary classifiers are trained on all class pairs of training data. An unknown pixel l is classified by applying all binary classifiers to its reflectance vector $\underset{\sim}{x}_l$ and then choosing the label ω_c, that is, setting $\hat{y}_l = c$, that is most frequently predicted (Karatzoglou, Meyer and Hornik, 2006).

The second voting scheme has been shown to work well with SVMs and, despite there being a larger number of SVMs to train, the overall CPU time used is less compared to the one-against-all voting method since the training datasets for each quadratic optimisation problem, which scales super-linearly, are smaller (Karatzoglou *et al.*, 2006).

There is limited research in the literature on how to statistically quantify the bias and variance properties of SVM predictions in the context of classification problems, as we wish to do for our crop area estimator, $\hat{T}_{cg}$, in (4.4) and (4.5), respectively, with respect to the underlying distributions of the crop classes. This presents a major challenge for us to put SVM techniques on a firm statistical inferential basis while at the same time preserving their computation advantages, as it is imperative that we provide objective statistical quality measures alongside our published official crop area estimates.

There are, however, a couple of options that we are aware of in the machine learning literature for statistically quantifying the accuracy of SVM results that we briefly discuss in Appendix D. Further evaluation of these methods is required in order to determine their suitability in appropriately estimating $\mathcal{B}\left(\hat{T}_{cg}\right)$ and $\mathrm{Var}\left(\hat{T}_{cg}\right)$ and we suspect that there are further approaches available in the literature to investigate beyond those mentioned in Appendix D.

5.2 Gaussian maximum likelihood classification

Gaussian maximum likelihood (ML) classification is one of the most widely used supervised classification techniques with satellite imagery data (Richards, 2013). It is not to be confused with maximum likelihood estimation. Gaussian ML classification is a parametric classification algorithm that assumes the data points in each class are generated from a multivariate normal distribution. Once the parameters of these distributions are estimated, based on the training dataset, labels for unknown pixels can be predicted by computing the probabilities of them being a member of each candidate class and assigning them the class with the highest probability.

If we knew the likelihood of ω_c being the correct label for an unknown pixel l given its position $\underset{\sim}{x}_l$ in the spectral space, for all $\omega_c \in \Omega$, that is, if we knew,

$$p_{c|\underset{\sim}{x}}\left(\underset{\sim}{x}_l\right) = \Pr\left(y_l = c \mid \underset{\sim}{x}_l\right), \quad c = 1,\ldots,C \,,$$

then we could classify the pixel according to the *Bayes decision rule*,

$$\underset{\sim}{x}_l \in \omega_c \ \text{ if } \ p_{c|\underset{\sim}{x}}\left(\underset{\sim}{x}_l\right) > p_{c'|\underset{\sim}{x}}\left(\underset{\sim}{x}_l\right) \ \text{ for all } \ c' \neq c \,. \tag{5.7}$$

These probabilities, unfortunately, are unknown but can be related to the class conditional probabilities, $p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right) = \Pr\left(\underset{\sim}{x}_l \mid y_l = c\right)$, and the *a priori* probabilities, $p_c = \Pr\left(y_l = c\right)$, which are able to be estimated, through Bayes theorem,

$$p_{c|\underset{\sim}{x}} = \frac{p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right) p_c}{\sum_{k=1}^{C} p_{\underset{\sim}{x}|k}\left(\underset{\sim}{x}_l\right) p_k} = \frac{p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right) p_c}{p_{\underset{\sim}{x}}\left(\underset{\sim}{x}_l\right)}, \tag{5.8}$$

where $p_{\underset{\sim}{x}}\left(\underset{\sim}{x}_l\right) = \Pr\left(\underset{\sim}{x}_l\right)$, the probability of finding a pixel with reflectance vector, $\underset{\sim}{x}_l$, in the image, from any class.

The class conditional probabilities, $p_{x|c}\left(\underset{\sim}{x}_l\right)$, describe the probability of finding a pixel at position $\underset{\sim}{x}_l$ in spectral space belonging to class, ω_c, which can be estimated from the training dataset, $\underset{\sim}{\mathcal{T}}$. As mentioned, it is assumed that data points belonging to crop class ω_c are generated from a multivariate normal probability distribution, $\text{Normal}\left(\underset{\sim}{\mu}_c, \mathbf{\Sigma}_c\right)$, where $\underset{\sim}{\mu}_c$ and $\mathbf{\Sigma}_c$ are an m dimensional mean vector and $m \times m$ dimensional covariance matrix for class ω_c, respectively, as it is generally appropriate for common spectral response distributions (Lillesand, Kiefer and Chipman, 2008). It is not entirely unreasonable to assume normality, as it would be expected that most data points in a distinct cluster would lie around the centre and would decrease in likelihood for positions further away from the centre where data points are less typical, and prediction accuracy is not overly sensitive to the normality assumption being violated (Richards, 2013).

The *a priori* probabilities, p_c, express the probability of class c occurring in the image of interest. These may be estimated from other sources of information, including ground surveys, existing maps and historical data (Schowengerdt, 2007); in our situation, we could use the ABS Agricultural Census (ABS, 2013) to inform these prior probabilities. If we're unsure of the suitability of the data we're using to inform these priors, we can assume them to be all equal.

As expressed via Bayes theorem in (5.8), the posterior probabilities for an unknown pixel l, $p_{c|\underset{\sim}{x}}\left(\underset{\sim}{x}_l\right)$, are proportional to its class conditional probabilities that are weighted by the priors. The $p_{\underset{\sim}{x}}\left(\underset{\sim}{x}_l\right)$ can be ignored as they are not class dependent and add no extra information for making the decision in (5.7).

It is mathematically convenient at this point to define the *discriminant function* for unknown pixel l belonging to crop ω_c, $g_c\left(\underset{\sim}{x}_l\right)$, as,

$$g_c\left(\underset{\sim}{x}_l\right) = \ln\left\{p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right)p_c\right\} = \ln p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right) + \ln p_c \ ,$$

which, if we assume our p_c are all equal (that is, flat priors), effectively becomes,

$$g_c\left(\underset{\sim}{x}_l\right) = -\ln\left|\Sigma_c\right| - \left(\underset{\sim}{x}_l - \underset{\sim}{\mu}_c\right)^{\mathrm{T}} \Sigma_c^{-1}\left(\underset{\sim}{x}_l - \underset{\sim}{\mu}_c\right),$$

for the *Gaussian ML classifier*, so that we can extend the decision rule in (5.7) to,

$$\underset{\sim}{x}_l \in \omega_c \ \text{ if } g_c\left(\underset{\sim}{x}_l\right) > g_{c'}\left(\underset{\sim}{x}_l\right) \ \text{ for all } c' \neq c \ . \qquad (5.9)$$

As mentioned earlier, we will most likely have prior knowledge about the *a priori* class probabilities from past ABS agricultural censuses and surveys, but for ease of exposition, we will assume flat priors for the remainder of the paper.

Equivalently, the decision rule in (5.9) for the Gaussian ML classifier can be expressed in the form of our general classifier function for $\hat{y}_l$, $\hat{\mathcal{F}}$, as defined in (4.3),

$$\mathcal{F}\left(\underset{\sim}{x}_l; \underset{\sim}{\mu}_c, \mathbf{\Sigma}_c : c = 1,\ldots,C\right) = \underset{c=1,\ldots,C}{\arg\max}\left\{g_c\left(\underset{\sim}{x}_l\right)\right\}. \qquad (5.10)$$

We estimate the $\underset{\sim}{\mu}_c$ and $\mathbf{\Sigma}_c$ parameters based on the training dataset, which we assume has sufficient data points for each crop class in order to accurately estimate the $\underset{\sim}{\mu}_c$ and $\mathbf{\Sigma}_c$, for each $c = 1,\ldots,C$. Swain and Davis (1978) suggest $10m$ data points per class as a minimum, but as many as $100m$ per class being preferable. This has not been difficult for us to achieve in creating our first training datasets but we believe we are not capturing the full variability, both within and between crop types; our experience suggests that including in the training dataset pixel information for each crop class from different locations (that is, farms) and from different times during the crop growing cycle is just as important, if not more so, than the number of data points in the training dataset in order to have sufficient information to accurately estimate the probability distribution associated with each crop.

The implication of the decision rule in (5.7) is that any unknown pixel l will be classified as one of the candidate crop classes ω_c regardless of how small the actual conditional probability of membership in that class is. Thresholds can be set by the analyst and incorporated into the decision rule so that if the probability of the most likely class to which the unknown pixel belongs is below that threshold, the pixel can be labelled as ‘unclassified’ (Lillesand *et al.*,2008).

Another implication of ML classification is that,

$$\sum_{k=1}^{C} p_{\underset{\sim}{x}|k}\left(\underset{\sim}{x}_l\right) = 1 ,$$

is a condition that is not necessarily satisfied for an unknown pixel l. Even though we're assuming a probability distribution for each class, it is not guaranteed that the probabilities of pixel l belonging to each class will sum to one.

For example, one can imagine a simple situation where our training dataset contains only two crop classes, ω_c and $\omega_{c'}$, that have very similar and overlapping probability distributions. An unknown pixel l may have a reflectance vector $\underset{\sim}{x}_l$ such that $p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right) = 0.85$ and $p_{\underset{\sim}{x}|c'}\left(\underset{\sim}{x}_l\right) = 0.82$, resulting in,

$$\sum_{k=c,c'} p_{\underset{\sim}{x}|k}\left(\underset{\sim}{x}_l\right) = 1.67 > 1 .$$

Using the decision rule in (5.7), we would label pixel l as ω_c but there is a very high chance that the land pixel l represents contains crop $\omega_{c'}$. One can imagine a different situation where, again, our training dataset contains only two crop classes, ω_c and $\omega_{c'}$, and an unknown pixel l has a reflectance vector $\underset{\sim}{x}_l$ such that $p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right) = 0.85$ and $p_{\underset{\sim}{x}|c'}\left(\underset{\sim}{x}_l\right) = 0.02$, resulting in,

$$\sum_{k=c,c'} p_{\underset{\sim}{x}|k}\left(\underset{\sim}{x}_l\right) = 0.87 < 1 .$$

The outcome of the decision rule in (5.7) would be the same, that is, we would label pixel l as ω_c but, this time, we are much more confident that this is the correct label.

In order for our measures of bias and variance for our estimator $\hat{T}_{cgt}$ to properly reflect the different situations and levels of confidence in the crop class predictions for pixels, as demonstrated in our example above, we propose to normalise the conditional probabilities and instead use,

$$\mathcal{P}_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right) = \frac{p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right)}{\sum_{k=1}^{C} p_{\underset{\sim}{x}|k}\left(\underset{\sim}{x}_l\right)} , \tag{5.11}$$

rather than $p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right)$ in the decision rule and equivalent classifier function in (5.9) and (5.10), respectively. The sum to one condition is satisfied for the $\mathcal{P}_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right)$, that is $\sum_{k=1}^{C} \mathcal{P}_{\underset{\sim}{x}|k}\left(\underset{\sim}{x}_l\right) = 1$, and our definition of $\mathcal{P}_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right)$ is consistent with Bayes' theorem, which is commonly used in discriminant analysis (Breheny, 2011). The use of the $\mathcal{P}_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right)$ does not change the outcome of the decision rule in (5.9), nor the classifier function in (5.10).

The bias of our estimator, $\hat{T}_{cg}$, under the Gaussian ML classification method is then given by,

$$\begin{aligned}\mathcal{B}\left(\hat{T}_{cg}\right) &= \sum_{i=1}^{N_g} a_i \mathcal{B}\left(\hat{Z}_{ic}\right) \\ &= \sum_{i=1}^{N_g} a_i \left\{E\left(\hat{Z}_{ic}\right) - Z_{ic}\right\} \\ &= \sum_{i=1}^{N_g} a_i \left\{\mathcal{P}_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_i\right) - Z_{ic}\right\},\end{aligned} \tag{5.12}$$

where we arrive at the final line in (5.12) by definition of $\hat{Z}_{ic}$, that is,

$$E\left(\hat{Z}_{ic}\right) = 1 \times \mathcal{P}_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_i\right) + 0 \times \sum_{\substack{k=1 \\ k \neq c}}^{C} \mathcal{P}_{\underset{\sim}{x}|k}\left(\underset{\sim}{x}_i\right) = \mathcal{P}_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_i\right).$$

The bias will be small when, for all $i = 1, \ldots, N_g$,

(i) for $Z_{ic} = 1$, $\mathcal{P}_{\underset{\sim}{x}|c}\left(x_i\right) \rightarrow 1$ which is equivalent to $p_{\underset{\sim}{x}|c}\left(x_i\right) \rightarrow \sum_{k=1}^{C} p_{\underset{\sim}{x}|k}\left(\underset{\sim}{x}_i\right)$; and

(ii) for $Z_{ic} = 0$, $\mathcal{P}_{\underset{\sim}{x}|c}\left(x_i\right) \rightarrow 0$ which is equivalent to $p_{\underset{\sim}{x}|c}\left(x_i\right) \rightarrow 0$.

The bias will be large when, for all $i = 1, \ldots, N_g$,

(iii) for $Z_{ic} = 1$, $\mathcal{P}_{\underset{\sim}{x}|c}\left(x_i\right) \rightarrow 0$ which is equivalent to $p_{\underset{\sim}{x}|c}\left(x_i\right) \rightarrow 0$ or

$\sum_{\substack{k=1 \\ k \neq c}}^{C} p_{\underset{\sim}{x}|k}\left(\underset{\sim}{x}_i\right) \rightarrow (C-1) \Rightarrow p_{\underset{\sim}{x}|c'}\left(x_i\right) \rightarrow 1$ for all $c' \neq c$; and

(iv) for $Z_{ic} = 0$, $\mathcal{P}_{\underset{\sim}{x}|c}\left(x_i\right) \rightarrow 1$ which is equivalent to $p_{\underset{\sim}{x}|c}\left(x_i\right) \rightarrow \sum_{k=1}^{C} p_{\underset{\sim}{x}|k}\left(\underset{\sim}{x}_i\right)$.

The effects of conditions (ii) and (iii) on the bias of $\hat{T}_{cg}$ make intuitive sense. Conditions (i) and (iv) are not anticipated in practice and appear at first to not be desirable but can be explained. If $p_{\underset{\sim}{x}|c}\left(x_i\right) \rightarrow \sum_{k=1}^{C} p_{\underset{\sim}{x}|k}\left(\underset{\sim}{x}_i\right)$, then $p_{\underset{\sim}{x}|c}\left(x_i\right) > p_{\underset{\sim}{x}|c'}\left(x_i\right)$ for $c' \neq c$. This condition thus ensures that we will be accurate at estimating $\hat{Z}_{ic} = 1$ when $Z_{ic} = 1$ and $\hat{Z}_{ic} = 0$ when $Z_{ic} = 0$, which, if it were satisfied for many of the N_g pixels, would result in a small bias for our estimate.

The variance of our estimator, $\hat{T}_{cg}$, under the Gaussian ML classification method is given by,

$$\begin{aligned}Var\left(\hat{T}_{cg}\right) &= \sum_{i=1}^{N_g} a_i^2 Var\left(\hat{Z}_{ic}\right) \\ &= \sum_{i=1}^{N_g} a_i^2 \left\{E\left(\hat{Z}_{ic}^2\right) - E\left(\hat{Z}_{ic}\right)^2\right\} \\ &= \sum_{i=1}^{N_g} a_i^2 \mathcal{P}_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_i\right)\left\{1 - \mathcal{P}_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_i\right)\right\}.\end{aligned} \tag{5.13}$$

Note that $\operatorname{Cov}\left(\hat{Z}_{ic}, \hat{Z}_{jc}\right) = 0$ for all $i \neq j$ since each pixel is classified independently. The variance of our estimator will be small when we are confident that each i-th pixel, for all $i = 1, \ldots, N_g$, is either a member or not of crop class ω_c, regardless of the value of Z_{ic} which indicates its true membership. That is, when

(i) $\mathcal{P}_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_i\right) \to 1$, which is equivalent to $p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_i\right) \to \sum_{k=1}^{C} p_{\underset{\sim}{x}|k}\left(\underset{\sim}{x}_i\right)$; or

(ii) $\mathcal{P}_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_i\right) \to 0$, which is equivalent to $p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_i\right) \to 0$.

The variance of our estimator will be larger when we aren't confident about the membership of each i-th pixel, for all $i = 1, \ldots, N_g$, in crop class ω_c, regardless of the value of Z_{ic}; the most extreme case, when the variance is at its maximum, is when $\mathcal{P}_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_i\right) \to 1/2$.

5.3 Classification with kernel density estimation

Classification using kernel density estimation can be seen to be an extension of Gaussian ML classification where the probability distributions of the candidate crop classes are allowed to be semiparametric and hence more flexible. Although Lillesand *et al.* (2008) and Richards (2013) claim that the multivariate normal distribution assumption is valid as predictions are not overly sensitive to the assumption being violated in most satellite imagery applications, we believe it is worthwhile investigating kernel density estimation to test this assumption. While still avoiding overfitting, we anticipate that capturing any significant distinguishing features of the probability distributions of the different crop classes when estimating them could lead to more accurate classification performance.

Using notation similar to that employed in Wand and Jones (1993), the most general form of the multivariate kernel estimator for our $p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right)$ is given by,

$$\hat{h}_c\left(\underset{\sim}{x}_l; \mathrm{H}_c\right) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathrm{H}_c}\left(\underset{\sim}{x} - \underset{\sim}{x}_l\right) , \tag{5.14}$$

where K is an m variate kernel function that is a multivariate probability density, H_c is a symmetric, positive definite $m \times m$ dimensional bandwidth (not to be confused with spectral bandwidths introduced in Subsection 2.1) matrix that specifies the amount and direction of smoothing and $K_{\mathrm{H}_c}\left(\underset{\sim}{x}\right) = \left|\mathrm{H}_c\right|^{-1/2} K\left(\mathrm{H}_c^{-1/2}\underset{\sim}{x}\right)$. Recall that n is the size of the training dataset defined in Section 4.

The choice of the kernel function K is not as crucial to the accuracy of the kernel density estimator as the choice of the bandwidth matrix H_c. For mathematical convenience, we can use the standard multivariate normal kernel $K\left(\underset{\sim}{x}\right) = \left(2\pi\right)^{-m/2} \exp\left(-\underset{\sim}{x}^{\mathrm{T}}\underset{\sim}{x} / 2\right)$ but choosing H_c is not a straight forward process.

Bandwidth selection methodology for the multivariate case has received considerable attention and is still an active area of research in the literature; for example, Zougab, Adjabi and Kokonendji (2014) investigate Bayesian estimation of adaptive bandwidth matrices in the multivariate case, using the quadratic and entropy loss functions. We have used the np package (Hayfield and Racine, 2008) in the R computing environment (R Development Core Team, 2014) which has various data-driven bandwidth selection methods to choose from, to estimate our multivariate kernel estimator. Hayfield and Racine (2008) caution that the data-driven bandwidth selection methods are computationally demanding; greater computational expense is the price to be paid for the flexibility of semiparametric kernel density estimation compared to its parametric counterparts. Scott and Wand (1991) show another disadvantage in that as the number of dimensions increases, the accuracy of multivariate kernel density estimation deteriorates and it requires an increase in the size of the training dataset to maintain acceptable accuracy levels.

Putting these implementation challenges aside, despite their importance, and assuming we are able to compute a kernel estimator of $p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right)$ that is of sufficient accuracy and with acceptable computational cost, then our pixel classification problem and the bias and variance properties of our estimator $\hat{T}_{cg}$ are the same as for Gaussian ML classification, as outlined in Subsection 5.2. That is:

- the decision rule for classifying an unknown pixel l with reflectance vector $\underset{\sim}{x}_l$ is as given in (5.9), where $p_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right)$ is replaced by $\hat{h}_c\left(\underset{\sim}{x}_l;\mathbf{H}_c\right)$ in $g_c\left(\underset{\sim}{x}_l\right)$, and similarly for $g_{c'}\left(\underset{\sim}{x}_l\right)$; and
- the bias and variance properties of $\hat{T}_{cg}$ are as given in (5.12) and (5.13), respectively, where,

$$\mathcal{P}_{\underset{\sim}{x}|c}\left(\underset{\sim}{x}_l\right) = \frac{\hat{h}_c\left(\underset{\sim}{x}_l;\mathbf{H}_c\right)}{\sum_{k=1}^{C}\hat{h}_k\left(\underset{\sim}{x}_l;\mathbf{H}_k\right)} \quad .$$

It is of interest for us to see if kernel density estimation provides improvement over the Gaussian ML classification method that warrants the extra computational cost and need for a significantly larger training dataset.

5.4 Multinomial logistic regression

Multinomial logistic regression (MLR) is a well-established statistical classification method that generalises logistic regression to a multiclass problem (Agresti, 2002).

In MLR, we assume that the probability distribution of the response variable is given by the multinomial distribution. In our case,

$$Y_i \mid \underset{\sim}{X}_i \sim \text{Multinomial}\left(\underset{\sim}{\pi}_i, 1\right) \ , \tag{5.15}$$

where

- Y_i is a random variable that takes on the value c if the crop type growing on the land represented by the i-th pixel is ω_c;
- $\underset{\sim}{\pi}_i = \left(\pi_{i1},\ldots,\pi_{iC}\right)^{\mathrm{T}}$ is a C dimensional vector with elements $\pi_{ic} = p_{c|\underset{\sim}{x}}\left(\underset{\sim}{X}_i\right) = \Pr\left(Y_i = c \mid \underset{\sim}{X}_i\right)$, for $c = 1,\ldots,C$, which make up a discrete probability distribution; and
- $\underset{\sim}{X}_i$ is a $(m+1)$ dimensional random vector containing an intercept term and the m spectral reflectance measurements for the i-th pixel.

We then assume that the log-odds of each response follow a linear model,

$$\log\left(\frac{\pi_{ic}}{\pi_{iC}}\right) = \underset{\sim}{X}_i\underset{\sim}{\beta}_c + \varepsilon_i \quad , \tag{5.16}$$

where $\underset{\sim}{\beta}_c = \left(\beta_{c0},\beta_{c1},\ldots,\beta_{cm}\right)^{\mathrm{T}}$ is a $(m+1)$ dimensional vector of regression coefficients, for $c = 1,\ldots,C$ and $\varepsilon_i \sim \text{Normal}\left(0,\sigma_\varepsilon^2\right)$ are i.i.d. random error terms. Note that our i.i.d. assumption on the ε_i implies that there is no covariance between the error terms of different responses and that the error terms of responses from different crop classes are generated from a common distribution. We may need to consider and allow for the log-odds, $\log\left(\pi_{ic}/\pi_{iC}\right)$, to be non-linearly related to the reflectance measurements in $\underset{\sim}{X}_i$. That is, we may need to employ model selection procedures to select which higher order covariate terms and interactions are significant in order to improve the model fit and, hence, the predictive power of the model.

We have a choice to make as to how we estimate the parameters, $\underset{\sim}{\beta}_c$, for $c = 1,\ldots,C$, and σ_ε^2. We can estimate them by using maximum likelihood estimation via a Newton-Raphson procedure, the theory and implementation of which is detailed in Czepiel (2002). We also have the option of setting priors for the $\underset{\sim}{\beta}_c$ and σ_ε^2, and estimating the MLR model via Markov Chain Monte Carlo (MCMC) techniques. Given that we don't have prior knowledge about the model parameters, we can use non-informative priors. For example, we could set diffuse Gaussian priors for the $\underset{\sim}{\beta}_c$, that is,

$$\underset{\sim}{\beta}_c \sim \text{Normal}\left(\underset{\sim}{0},\sigma_{\beta_c}^2\mathbf{I}_{(m+1)}\right) \quad , \quad \text{for } c = 1,\ldots,C \ ,$$

with the values of the hyperparameters $\sigma_{\beta_c}^2$ chosen to be large enough, say $\sigma_{\beta_c}^2 = 10^8$, so that the normal distribution is more or less uniform over the range of $\underset{\sim}{\beta}_c$ (Marley and Wand, 2010), and we could set the prior for the standard deviation parameter σ_ε to be,

$$\sigma_\varepsilon \sim \text{Half-Cauchy}\left(A_\varepsilon\right),$$

with hyperparameter, $A_{\varepsilon} = 25$, as recommended by Gelman (2006) for achieving non-informativeness for variance parameters.

Regardless of the estimation method we employ to fit the MLR model, we will achieve estimates of the $\underset{\sim}{\beta}_c$ and σ_{ε}^2 , namely $\hat{\underset{\sim}{\beta}}_c$ and $\hat{\sigma}_{\varepsilon}^2$, respectively, based on the training dataset that will allow us to estimate the $\underset{\sim}{\pi}_i$, namely $\hat{\underset{\sim}{\pi}}_i$. That is, by rearranging (5.16) to get an expression for the π_{ic} in terms of the $\underset{\sim}{\beta}_c$, and substituting in $\hat{\underset{\sim}{\beta}}_c$, we can estimate the elements of $\hat{\underset{\sim}{\pi}}_i$ using,

$$\begin{aligned} \hat{\pi}_{ic} &= \frac{\exp\left(\underset{\sim}{X}_i \hat{\underset{\sim}{\beta}}_c\right)}{1 + \sum_{k=1}^{C-1} \exp\left(\underset{\sim}{X}_i \hat{\underset{\sim}{\beta}}_k\right)}, \quad \text{for } c = 1,\ldots,C-1, \quad \text{and} \\ \hat{\pi}_{iC} &= 1 - \sum_{k=1}^{C-1} \hat{\pi}_{ik} \ . \end{aligned} \tag{5.17}$$

Given (5.17), we can predict $\hat{\underset{\sim}{y}}_l$ for an unknown pixel l using the estimated assignment function,

$$\hat{y}_l = \hat{\mathcal{F}}\left(\underset{\sim}{x}_l; \hat{\mu}_c, \hat{\sigma}_{\varepsilon}^2 : c = 1,\ldots,C\right) = \underset{c=1,\ldots,C}{\arg\max}\left\{\hat{\pi}_{lc}\right\}. \tag{5.18}$$

Using (5.15), with $\hat{\underset{\sim}{\pi}}_i$ substituted in for $\underset{\sim}{\pi}_i$, and the known mean and variance formulae for the multinomial distribution, the bias of our estimator $\hat{T}_{cg}$ under the MLR model is given by,

$$\begin{aligned} \mathcal{B}\left(\hat{T}_{cg}\right) &= \sum_{i=1}^{N_g} a_i \mathcal{B}\left(\hat{Z}_{ic}\right) \\ &= \sum_{i=1}^{N_g} a_i \left\{E\left(\hat{Z}_{ic}\right) - Z_{ic}\right\} \\ &= \sum_{i=1}^{N_g} a_i \left\{\hat{\pi}_{ic} - Z_{ic}\right\}, \end{aligned}$$

and the variance of $\hat{T}_{cg}$ is given by,

$$\begin{aligned} Var\left(\hat{T}_{cg}\right) &= \sum_{i=1}^{N_g} a_i^2 Var\left(\hat{Z}_{ic}\right) \\ &= \sum_{i=1}^{N_g} a_i^2 \left\{E\left(\hat{Z}_{ic}^2\right) - E\left(\hat{Z}_{ic}\right)^2\right\} \\ &= \sum_{i=1}^{N_g} a_i^2 \hat{\pi}_{ic} \left\{1 - \hat{\pi}_{ic}\right\}. \end{aligned}$$

6. PROCESS FOR CREATING TRAINING AND TEST DATASETS

While other Australian organisations, including those discussed in Section 3, have been pursuing the classification of satellite data at the crop type level for decades, the lack of adequate ground-truth data has proved to be a major obstacle. The ABS is in the unique position of having unit record level data collected as part of the Rural Environment and Agricultural Statistics Program, in particular, the Agricultural Census (ABS, 2013). Although this data alone cannot be used as ground-truth data, when combined with address and property boundary data available from the PSMA Australia Limited (PSMA), it is a valuable source of reference information. There are, however, limitations and practical difficulties in creating quality training data using the Agricultural Census, the process for which we outline below.

Colleagues in the Geography Section of the ABS matched the addresses reported by respondents of the 2010/11 Agricultural Census (ABS, 2013) against the Geocoded National Address File (GNAF) from PSMA, and mapped the resulting spatial locations to PSMA land parcel boundaries. A dataset containing only high quality matches was passed on to us, consisting of 10.7% of the 135500 respondents to the 2010/11 Agricultural Census (ABS, 2012). The quality of address geocoding is, unfortunately, not consistent across Australia, and relatively higher in Victoria. Consequently, we have only sourced training data from Victoria so far which is likely to be introducing unknown levels of bias into the training dataset and not capturing enough variability.

To ensure accuracy in the training data, we need to know that a pixel selected within a property boundary contains the crop the respondent has reported. Thus the dataset containing the Victorian respondents with accurately geocoded addresses was subset further to those who reported growing a single crop type, only considering commonly grown cereal crops or similar, on most of the land contained within their property. Of the 50 respondents who reported growing a single crop type, only 15 reported that the crop they grew covered an adequate majority portion of their land. This limited the number of crop types in the training dataset to five.

For these 15 respondents, satellite images containing their properties were downloaded from the USGS Earth Explorer data service (USGS, 2014). We aimed to select images from the end of the growing season, as we assume that crops are most distinguishable when they are almost ready for harvest (an assumption which may not be valid and we wish to test in future investigations). Many images contained some level of cloud cover, rendering them unusable for us to extract pixels of particular crop types that weren't contaminated by clouds. Despite satellite images covering a particular area being available every 16 days, for a few candidate training data sites, it was not possible to find an image over a several month period for which the site was not covered by clouds.

For most of the respondents for which satellite images were downloaded, selecting pixels within the property boundary that we could label, with a relatively high degree of confidence, as the crop type reported by the respondent was a straightforward process. It was difficult for several properties, however, to distinguish which pixels were the crop of interest because either the 30m resolution did not allow for distinction between crops and other features or there appeared to be multiple blocks of different vegetation within the property boundary. We speculate that the obvious colour differences within the property boundary indicate different crop types, soil conditions or a partially harvested crop, but without further information we could not be sure which pixels to select and label as the crop reported by the respondent.

As a result of this process and the identified obstacles, our initial training dataset contained 1787 pixels in total for five crop types, selected from the farms of seven respondents in Victoria. Using this training dataset, we trained the four classifiers detailed in Section 5. In order to test their performance, we needed to create test datasets.

To create test datasets, we identified respondents, from those with accurately geocoded addresses, who grew one or more of the crop types included in the training dataset. Satellite images covering these respondents' properties, that contained no cloud cover and minimal missing data (due to the SLC failure problem, mentioned in Subsection 2.2) within the property boundary, were downloaded. This ensured that the reported areas of each crop type could be compared against our estimated areas of each crop type, generated from the pixel level predictions of each trained classifier, to assess the accuracy and quality of our classification methods.

We are aware that out current training dataset is very limited, restricting our ability to properly evaluate our classification methods. Work is currently underway to improve the training dataset by:

- including pixels from candidate training data sites in other states;
- broadening our criteria for candidate training data sites to include respondents who report growing more than one crop type on most of the land contained within their property and then implementing unsupervised clustering techniques to identify pixels from within their property boundary as the different crop types they reported; and
- in the longer term, accurately geocoding the addresses of all respondents on the Agricultural Census so that more training data sites are available to us (our Geography Section colleagues are conducting this work).

7. CONCLUSIONS AND LONG-TERM RESEARCH DIRECTIONS

7.1 Concluding remarks

The ABS' research efforts into the feasibility of using satellite imagery data to partially replace, supplement and validate data collected via the traditional surveys and censuses of the ABS Rural Environment and Agricultural Statistics Program have been underway for approximately six months. This paper is the result of our rapid learning and initial experiences in handling and analysing satellite imagery data during this period, where our focus has been on alternative methods for producing crop area estimates that accurately classify satellite imagery pixels to crop types.

We have not included in this paper our initial results from training the classification methods we detailed in Section 5, using the training data we describe in Section 6, and applying the trained classifiers to our test datasets, which are also described in Section 6. The reason we haven't included these initial results is due to our reservations about the quality of our training data, given the practical difficulties we faced in creating the training data, the unknown levels of bias our training data creation process may have introduced and the probable lack of representativeness of the training data with respect to the variability of reflectance measurements within and between different crop classes (see Section 6). With training data of questionable quality being used to train the classifiers, it follows that the resulting predictions and crop area estimates are questionable as well. While a few of our crop area estimates quite accurately matched the corresponding area reported by the respondents to the Agricultural Census 2010/11, many more were considerably off-target. We have yet to investigate why our few accurate crop area estimates were on-target.

Clearly, the practical issues we face in creating training data and test datasets of sufficient quality are the major obstacles for our research to continue to progress, as this is what we require to properly assess the classification methods we have focussed on in Section 5, and also, further methodologies and approaches we wish to investigate in the future. As mentioned in Section 6, work is currently being undertaken to improve the training datasets.

Although there is great interest among NSOs to realise the full potential of satellite imagery data in the production of official agricultural statistics, methodologies for estimating reliable agricultural statistics and their associated errors using satellite imagery data don't seem to exist yet in the literature. Other NSOs are concentrating their R&D effort on the problem of reliably forecasting crop yields using satellite imagery data. Even though we would like to eventually achieve this as well, we believe R&D effort into methods for reliably estimating crop area statistics is a better starting point, since crop yield is the result of a much more complicated real world process that is dependent on many factors including climate, weather and soil conditions, plant health and insecticide usage; data for which would need to be incorporated into crop yield forecasting models.

Our proposed formulation of land area estimation based on the classification of satellite imagery data, detailed in Section 4, appears to be a novel approach for creating official crop area statistics. It is a very interesting estimation problem, with many facets to address (see Subsection 7.2), that has the potential to pave the way for the production of new and richer agricultural statistical products to be possible in the future. Despite there being many challenges to overcome to achieve our goal of reliably estimating crop area statistics from satellite imagery data, we hope that these challenges will become increasingly easier to overcome over time through the sharing of methods and practices with other NSOs and other relevant research bodies, and the continual advancement of relevant technologies that will allow richer data sources to become available. For example, the Soil Moisture Active Passive (SMAP) satellite, that has been designed to collect data on soil moisture, is being launched by NASA in October 2014 (NASA, 2014). The data the SMAP satellite will collect has the potential to improve the crop yield forecasting models being developed by various NSOs.

7.2 Future research directions

Given that we are in the infancy of our satellite imagery data research program, we have approached the crop area estimation problem by first investigating methods for handling a simplified version of the estimation problem, stripped of many of its complications. The intention is that we will build a basic method to solve this simplified version of the estimation problem, which will then provide a solid foundation for the addition of further layers of complexity that will deal with methodological issues such as adjusting for cloud cover and pixels containing multiple land cover types (see Appendix A).

There are further lines of research for us to pursue, even for the simplified version of the estimation problem we are currently working to address, including:

Classifiers in the statistical literature

> We are aware that there are numerous classification methods in the statistical literature that we can apply to our satellite data classification problem, such as the linear discriminant and quadratic discriminant classifiers; see Delaigle and Hall (2013). We would appreciate the Committee's advice as to which statistical classifiers perform well, in terms of both prediction accuracy and computational efficiency, on high-dimensional data that are likely to belong to similar and overlapping class distributions, and also consider all possible classes simultaneously when solving a multi-class classification problem.

Establishing a statistical inferential basis for Support Vector Machines

As we have previously mentioned, the computational efficiency and prediction accuracy of SVMs for high dimensional data classification problems has led to their extensive use in satellite imagery analysis literature and applications, but in order for us to use SVMs to produce official crop area statistics, we need to develop a sound statistical inferential basis upon which to assess their performance. Appendix D highlights some of the machine learning literature that attempts to build a statistical inferential foundation for SVMs. The proposed approaches require further critical examination as, at first glance, they don't appear to be appropriate for our needs.

Supervised and unsupervised hybrid approaches to classification

Hybrid supervised and unsupervised classification techniques are often used to classify satellite imagery data as the splitting of the data into spectral classes via unsupervised classification methods can improve the performance of a subsequently applied supervised classifier (Richards, 2013).

Once we have solved the simplified version of the estimation problem to within acceptable levels of reliability and quality, we hope to eventually build a sophisticated, overall methodological routine for estimating official crop area statistics for Australia that handles all the complexities of the estimation problem. This overall methodological routine may require the incorporation of the following, which will shape our long-term research directions:

Classification with spatio-temporal models

As mentioned in Subsection 2.2, the supervised classification methods we have focussed on so far only consider the spectral reflectance measurements of the image pixels at a particular point in time; but the spectral reflectance measurements of that pixel in nearby time points and also the spatial context of the pixel would provide a lot of valuable information for the classification problem. Classification methods based on spatio-temporal models, which are examined extensively in Cressie and Wikle (2011), that account for the spatial and temporal relationships between pixels' reflectance measurements have the potential to be more accurate than classification methods that predict the correct label for each pixel independently, based only on the reflectance measurements of that pixel.

Imputation methods for missing data

Cloud cover is a major source of missing data in satellite imagery and could prove to be one of the greatest challenges to overcome. Certain crop types may be more susceptible to cloud cover due to the crop's requirement for higher levels of rainfall. For example, field peas are more likely to be grown in dairy producing areas with corresponding higher rainfall levels. Crops grown in the tropical areas of Australia are likely to have missing values due to cloud cover during the wet season. Other crops, such as wheat, are more commonly grown in drier, arid areas with less rainfall and, hence, may be less likely to be affected by cloud cover. Cloud cover, thus, can potentially lead to a bias in how various crop types are represented in training datasets, which could in turn result in a bias in particular crop area statistics. Appropriate imputation methods will need to be established to mitigate this risk of bias. As mentioned in Subsection 2.2, failures in on board satellite equipment are another potential source of vast amounts of missing data that may require different imputation treatments.

Measurement error models

Thin cloud cover is difficult to detect and can contaminate the reflectance measurements of the affected pixels. The reflectance measurements of pixels can also be contaminated by the shadows of clouds and high topographical land features. The development of methods to detect such contamination and the employment of classification methods based on measurement error models are a possible solution to this problem.

Unmixing methods

As demonstrated in Appendix A, the 30m resolution of Landsat-7 images leads to many pixels containing multiple types of land cover. Not accounting for this within-pixel *mixing* will affect the accuracy of the classifiers applied to the satellite imagery data. Plaza, Du, Bioucas-Dias, Jia and Kruse (2011) provides a good range of techniques for spectral unmixing of satellite imagery data.

Sample design theory for training data selection

The practical difficulties we face in creating training data determine our training data selection method; we raised our concerns in Section 6 about the potential but unknown biases our current training data creation process introduces into, and the insufficient amount of variability of reflectance measurements within and between different crop classes it captures in the resulting training dataset. As recommended in Lillesand *et al.* (2008), it is preferable to gather training data using sampling design principles appropriate for the particular application at hand. We hope to investigate this further once the more fundamental training data creation issues have been addressed.

Optimisation of regional classifiers

Given the size of Australia and its vast climatic and soil conditions, different areas of Australia are appropriate for growing different crops. It would not be practical to apply a global classifier, that considers all crops grown in Australia, to all pixels within the agricultural land areas of Australia. A much better approach that would conceivably be more efficient and accurate would be to develop multiple regional classifiers. Determining the regions that optimise the overall efficiency and accuracy of all the regional classifiers would be a difficult problem to solve. Generating training datasets for each regional classifier would be another aspect to cover, as would determining the frequency with which the regions would need to be revised and the regional classifiers recalibrated, or re-trained. Regional agronomic data, which is currently limited in Australia, could be incorporated into such a system of regional classifiers, if it were to become available in the future.

Hyperspectral data

We only have multispectral satellite imagery data available to us currently but with continual advancements in relevant technologies, it is possible that hyperspectral satellite imagery data will be available in the future. Classifying such high dimensional data is computationally infeasible for most statistical classifiers, which will mean we will need to research dimension reduction methods, called feature selection methods in the satellite imagery literature, and assess their statistical merits. A general overview of feature selection methods is given in Jia *et al.* (2013).

REFERENCES

Agresti, A. (2002) *Categorical Data Analysis*, Second Edition, Wiley, New York.

Australian Bureau of Agricultural and Resource Economics and Sciences (2011) *Guidelines for Land Use Mapping in Australia: Principles, Procedures and Defintions*, Fourth Edition, ABARES, Canberra.

Australian Bureau of Statistics (2011) *Australian Statistical Geography Standard (ASGS): Volume 3 – Non ABS Structures, July 2011*, cat. no. 1270.0.55.003, ABS, Canberra.

—— (2012) *Agricultural Commodities, Australia, 2010–11*, cat. no. 7120.0, ABS, Canberra.

—— (2013) *2011 Agricultural Census*, webpage.
< http://www.abs.gov.au/websitedbs/c311215.nsf/web/Agriculture+-+Agricultural+Census >

Bedard, F. and Reichert, G. (2013) "The Use of Remote Sensing for Agricultural Statistics: Preliminary Results and Findings", Internal Report, Statistics Canada; Ottawa, Canada.

Ben-Hur, A. and Weston, J. (2009) "A User's Guide to Support Vector Machines" in *Biological Data Mining*, Springer Protocols.

Berliner, L. (1996) "Hierarchical Bayesian Time Series Models", in *Maximum Entropy and Bayesian Methods*, Kluwer Academic Publishers, Dordrecht, Netherlands.

Boryan, C.; Yang, Z.; Mueller, R. and Craig, M. (2011) "Monitoring U.S. Agriculture: The U.S. Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program", *Geocarto International*, 26(5), pp. 341–358.

Boser, B.; Guyon, I. and Vapnik, V. (1992) "A Training Algorithm for Optimal Margin Classifiers", in *Proceedings of the Fifth Annual Workshop of Computational Learning Theory*, 5, pp. 144–152, Pittsburgh, ACM.

Breheny, P. (2011) "Linear Discriminant Analysis, Part I", lecture slides for *Applied Statistical Modelling*, University of Kentucky, accessed online.
< http://web.as.uky.edu/statistics/users/pbreheny/764-F11/notes/9-15.pdf >

Commonwealth Science and Industrial Research Organisation (2011) *Environmental Earth Observation*, webpage.
< http://www.csiro.au/Organisation-Structure/Divisions/Land-and-Water/Environmental-Earth-Observation.aspx >

—— (2011) *Atmosphere and Land Observation and Assessment*, webpage.
< http://www.csiro.au/en/Organisation-Structure/Divisions/Marine--Atmospheric-Research/Atmosphere-land-observation.aspx >

Congalton, R. and Green, K. (2009) *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Second Edition, CRC Press, Boca Raton.

Cortes, C. and Vapnik, V. (1995) "Support Vector Networks", *Machine Learning*, 20, pp. 273–297.

Crammer, K. and Singer, Y. (2002) "On the Learnability and Design of Output Codes for Multiclass Problems", *Machine Learning*, 47, pp. 201–233.

Cressie, N. and Wikle, C. (2011) *Statistics for Spatio-Temporal Data*, Wiley, New York.

Czepiel, S. (2002) *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*, accessed online.
< http://czep.net/stat/mlelr.pdf >

Delaigle, A. and Hall, P. (2013) "Classification Using Censored Functional Data", *Journal of the American Statistical Association*, 108 (504), pp. 1269-1283.

Gelman, A. (2006) "Prior Distributions for Variance Parameters in Hierarchical Models", *Bayesian Analysis*, 1, pp. 515–533.

Gualtieri, J. and Cromp, R. (1998) "Support Vector Machines for Hyperspectral Remote Sensing Classification", in *Proceedings of the 27-th AIRP Workshop: Advances in Computer Assisted Recognition*, pp. 221–232, Washington, D.C.

Han, W.; Yang, Z.; Di, L. and Mueller, R. (2012) "CropScape: A Web Service Based Application for Exploring and Disseminating U.S. Conterminous Geospatial Cropland Data Products for Decision Support", *Computers and Electronics in Agriculture*, 84, pp. 111–123.

Hayfield, T. and Racine, J. (2008) "Nonparametric Econometrics: The np Package", *Journal of Statistical Software*, 27(5), pp. 1–32.
< http://www.jstatsoft.org/v27/i05/ >

Henderson, A. and Pitchford, S. (2013) "ABS' Rural Environment and Agricultural Statistical Collections into the Future", presented at the *Sixth International Conference on Agricultural Statistics*, Rio de Janeiro, Brazil.

Hsu, C.-W.; Chang, C.-C. and Lin, C.-J. (2010) *A Practical Guide to Support Vector Classification*, Technical Report, Department of Computer Science, National Taiwan University.

Jia, X.; Kuo, B. and Crawford, M. (2013) "Feature Mining for Hyperspectral Image Classification", *Proceedings of the IEEE*, 101(3), pp. 676–697.

Jones, H. and Vaughan, R. (2010) *Remote Sensing of Vegetation: Principles, Techniques and Applications*, Oxford University Press, New York.

Karatzoglou, A.; Meyer, D. and Hornik, K. (2006) “Support Vector Machines in R”, *Journal of Statistical Software*, 15(9), pp. 1–28.

Kruppa, J.; Liu, Y.; Biau, G.; Kohler, M.; Koenig, I.; Malley, J. and Ziegler, A. (2014) “Probability Estimation with Machine Learning Methods for Dichotomous and Multicategory Outcome: Theory”, *Biometrical Journal* (to appear).

Lee, Y.; Lin, Y. and Wahba, G. (2004) “Multicategory Support Vector Machines: Theory and Application to the Classification of Microarray Data and Satellite Radiance Data”, *Journal of the American Statistical Association*, 99(465), pp. 67–81.

Lillesand, T.; Kiefer, R. and Chipman, J. (2008) *Remote Sensing and Image Interpretation*, Sixth Edition, Wiley, New York.

Lin, H.; Lin, C. and Weng, R. (2007) “A Note on Platt’s Probabilistic Outputs for Support Vector Machines”, *Machine Learning*, 68(3), pp. 267–276.

Liu, W.; Gopal, S. and Woodcock, C. (2004) “Uncertainty and Confidence in Land Cover Classification Using a Hybrid Classifier Approach”, *Photogrammetric Engineering & Remote Sensing*, 70(8), pp. 963–971.

Lymburner, L.; Tan, P.; Mueller, N.; Thackway, R.; Lewis, A.; Thankappan, M.; Randall, L.; Islam, A. and Senarath, U. (2011) *The National Dynamic Land Cover Dataset – Technical Report*, Geoscience Australia & Australian Bureau of Agricultural and Resource Economics and Sciences; Canberra.

Marley, J. and Wand, M. (2010) “Non-Standard Semiparametric Regression via BRugs”, *Journal of Statistical Software*, 37(5), pp.1–30.
< http://www.jstatsoft.org/v37/i05/ >

Mentch, L. and Hooker, G. (2014) *Ensemble Trees and CLTs: Statistical Inference for Supervised Learning*, Cornwell University Library, New York.
(arXiv:1404.6473)

Meurink, A. (2013) *Estimating Arable Crop Yields Through Satellite Imaging*, Internal Report, Centraal Bureau voor de Statistiek; Den Haag, Netherlands.

National Aeronautics and Space Administration (2014) *Landsat Science – Landsat 8*, online information page.
< http://landsat.gsfc.nasa.gov/?p=3186 >

——(2014) *Soil Moisture Active Passive – Mission Imperative*, online information page.
< https://smap.jpl.nasa.gov/Imperative/ >

National Computational Infrastructure (2013) *Data Cube – The Future of Earth Observation*, online news article.
< http://nci.org.au/2013/12/11/data-cube-future-earth-observation/ >

Plaza, A.; Du, Q.; Bioucas-Dias, J.; Jia, X. and Kruse, F. (eds.) (2011) "Special Issue on Spectral Unmixing of Remotely Sensed Data", *IEEE Transactions on Geoscience and Remote Sensing*, 49 (11).

R Development Core Team (2014) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna.
< http://www.R-project.org/ >

Richards, J. (2013) *Remote Sensing Digital Image Analysis*, Springer, Berlin.

Schowengerdt, R. (2007) *Remote Sensing: Models and Methods for Image Processing*, Third Edition, Academic Press, San Diego.

Scott, D. and Wand, M. (1991) "Feasibility of Multivariate Density Estimates", *Biometrika*, 78, pp. 197–206.

Stewart, J.; Rickards, J. and Randall, L. (2013) *Ground Cover Monitoring for Australia: Progress Report, July 2011 to July 2012*, ABARES Technical Report 13.5, Australian Bureau of Agricultural and Resource Economics and Sciences, Canberra.

Swain, P. and Davis, S. (eds.) (1978) *Remote Sensing: The Quantitative Approach*, McGraw-Hill, New York.

Tam, S.-M. and Clarke, F. (2014) "Small Steps Towards Big Data: Some Initiatives by the Australian Bureau of Statistics", submitted to *International Statistical Review*.

Terrestrial Ecosystem Research Network (2012) "Using Google to Deliver Spatial Science", *TERN e-Newsletter*.
< http://www.tern.org.au/Newsletter-2012-Nov-GoogleEarthEngine-pg24291.html >

—— (2012) "NCRIS Partners Work Together to Build Soils-to-Satellites Tool", *TERN e-Newsletter*.
< http://www.tern.org.au/Newsletter-2012-Aug-NCRISSoils2Satellites-pg23207.html >

U.S. Geological Survey (2014) *EarthExplorer Data Service*.
< http://earthexplorer.usgs.gov/ >

——(2013) *SLC-Off Products: Background*, online information page.
< http://landsat.usgs.gov/products_slcoffbackground.php >

—— (2013) *Landsat – A Global Land-Imaging Mission*, U.S. Geological Survey; Sioux Falls, South Dakota, accessed online.
< http://pubs.usgs.gov/fs/2012/3072/fs2012-3072.pdf >

Wand, M. and Jones, M. (1993) "Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation", *Journal of the American Statistical Association*, 88(422), pp. 520–528.

Wu, T.; Lin, C. and Weng, R. (2004) "Probability Estimates for Multi-class Classification by Pairwise Coupling", *Journal of Machine Learning Research*, 5, pp. 975–1005.

Zhang, F.; Zhu, Z.; Pan, Y.; Hu, T. and Zhang, J. (2010) "Application of Remote Sensing Technology in Crop Acreage and Yield Statistical Survey in China", *Meeting on the Management of Statistical Information Systems (MSIS 2010)*, WP. 16.

Zhang, Z. and Jordan, M. (2006) "Bayesian Multicategory Support Vector Machines", in *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence.*

Zougab, N.; Adjabi, S. and Kokonendji, C. (2014) "Bayesian Estimation of Adaptive Bandwidth Matrices in Multivariate Kernel Density Estimation", *Computational Statistics & Data Analysis*, 75, pp. 28–38.

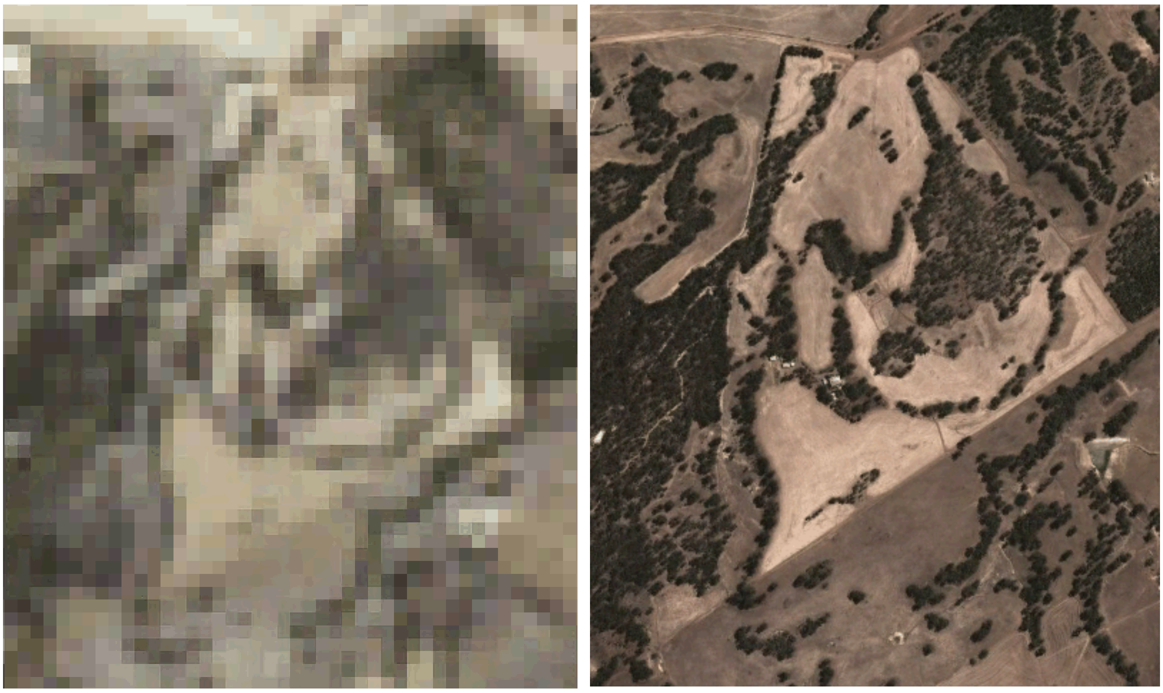
All URLs last viewed on 4 September 2014

APPENDIXES

A. EXAMPLE OF SATELLITE IMAGERY PIXELS CONTAINING MULTIPLE GROUND COVER TYPES

The images in figure A.1 demonstrate that the pixels in Landsat-7 images, that represent a 30m × 30m area of land, can contain multiple ground cover types. The image on the left is a visualisation of Landsat-7 satellite imagery data for a particular area of land in Australia, courtesy of the U.S. Geological Survey (USGS). The image on the right, covering the same land area, is courtesy of Google Earth. While the distinction between trees, crops and other ground cover types is clearly visible in the Google Earth image, the definition in the USGS image is considerably less, causing it to be blurred with many pixels containing combinations of trees, crops, bare soil, roads and/or other man-made structures.

A.1 An example of satellite imagery pixels containing multiple ground cover types



The image on the left, a visualisation of satellite imagery data courtesy of the U.S. Geological Survey, is a lower resolution version of the image on the right, courtesy of Google Earth (Map data: Google, Digital Globe, 2014). On comparing the two images, it is clear that many of the pixels in the image on the left contain multiple ground cover types.

B. EXTENSIONS TO THE LINEAR SUPPORT VECTOR MACHINE

We assumed in Subsection 5.1 for our binary SVM classifier case that the training data points for the two classes were linearly separable and so (5.5) classifies each of the training data points correctly. It is not likely in practice, however, that our two classes of training data points will be linearly separable but instead will exhibit some kind of overlap. To overcome this situation, we can allow for a larger margin and also allow the classifier function to misclassify some of the training data points by fitting a *soft-margin* SVM (Cortes *et al.*, 1995) which requires the quadratic optimisation problem in (5.3) to be reworked as

$$\begin{aligned} &\underset{\underset{\sim}{w},b}{\text{minimise}} \quad \frac{1}{2}\|\underset{\sim}{w}\|^2 + \mathcal{C}\sum\nolimits_{i=1}^{n}\xi_i \\ &\text{subject to:} \quad y_i(\underset{\sim}{w}^{\mathrm{T}}\underset{\sim}{x}_i + b) \geq 1 - \xi_i \quad i = 1,\ldots,n\,, \end{aligned} \tag{B.1}$$

where $\xi_i \geq 0$ are *slack variables* that let a training data point to be inside the margin $\left(0 \leq \xi_i \leq 1\right)$ or to be misclassified ($\xi_i > 1$), and the constant $\mathcal{C} > 0$ is a cost parameter that sets the importance of maximising the margin and minimising the amount of slack (Ben-Hur *et al.*, 2009). Again, by way of Lagrange multipliers, the dual representation of (B.1) is given by (5.4) but where the α_i are constrained by $0 \leq \alpha_i \leq \mathcal{C}$.

We stated in Subsection 5.1 that linear SVMs only depend on the input training data through dot products, as can be seen explicitly in (5.4). If we substitute a non-linear kernel function, $k\left(\underset{\sim}{x}_i, \underset{\sim}{x}_j\right)$, for the dot products, $\underset{\sim}{x}_i^{\mathrm{T}}\underset{\sim}{x}_j$, in (5.4), then we can find an optimal non-linear decision surface for the case when the training data points for the two classes are non-linearly separable. The dual representation of the quadratic optimisation problem for a fully generalised binary SVM classifier that allows for an optimal non-linear decision surface with a soft margin to be found is thus given by

$$\begin{aligned} &\underset{\underset{\sim}{\alpha}}{\text{maximise}} \quad \sum\nolimits_{i=1}^{n}\alpha_i - \frac{1}{2}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{n}\alpha_i\alpha_j u_i u_j \underset{\sim}{x}_i^{\mathrm{T}}\underset{\sim}{x}_j \\ &\text{subject to:} \quad \begin{cases} \text{i.} \quad \sum\nolimits_{i=1}^{n}\alpha_i u_i = 0\,, \\ \text{ii.} \quad \alpha_i \geq 0\,, \\ \text{iii.} \quad \alpha_i\left\{u_i\left(\underset{\sim}{w}^{\mathrm{T}}\underset{\sim}{x}_i + b\right) - 1\right\} = 0, \quad i = 1,\ldots,n\,. \end{cases} \end{aligned} \tag{B.2}$$

The use of the ***kernel trick*** is only worthwhile if the kernel function, defined as $k\left(\underset{\sim}{x}_i, \underset{\sim}{x}_j\right) = \phi\left(\underset{\sim}{x}_i\right)^{\mathrm{T}}\phi\left(\underset{\sim}{x}_j\right)$ where $\phi : \mathbb{R}^m \to \mathbb{R}^{\mathcal{H}}$ and $\mathcal{H} > m$, can be computed efficiently, that is, the mapping ϕ does not need to be explicitly computed for each $\underset{\sim}{x}_i$ in the training set.

A popular kernel function in satellite imagery classification applications is the Gaussian radial basis function (RBF) kernel (Richards, 2013), given by

$$k\left(\underset{\sim}{x}_i,\underset{\sim}{x}_j\right)=\exp\left\{-\gamma\left\|\underset{\sim}{x}_i-\underset{\sim}{x}_j\right\|^2\right\},\quad \gamma>0\,.$$

The linear kernel is a special case of the RBF given certain parameter combinations and so an SVM will fit a linear kernel if it is appropriate for the problem at hand, making the RBF a reasonable first choice (Hsu, Chang and Lin, 2010). Other common kernel functions include:

- Polynomial: $k\left(\underset{\sim}{x}_i,\underset{\sim}{x}_j\right)=\gamma\left(\underset{\sim}{x}_i^{\mathrm{T}}\underset{\sim}{x}_j+a_0\right)^d$; and
- Sigmoid: $k\left(\underset{\sim}{x}_i,\underset{\sim}{x}_j\right)=\tanh\left\{\gamma\underset{\sim}{x}_i^{\mathrm{T}}\underset{\sim}{x}_j+a_0\right\}$.

C. MULTI-CLASS SUPPORT VECTOR MACHINES

Rather than extending the binary SVMs to the multi-class problem, Crammer and Singer (2002) reformulated the support vector quadratic optimisation problem to handle multiple classes and proposed an algorithm that works by solving a single optimisation problem using data for all the classes.

There have been a number of other proposed multi-class SVMs but Zhang and Jordan (2012) consider the work of Lee, Lin and Wahba (2004) to be the most principled approach. Their approach considers the multiple classes jointly through an optimal classification rule. They define $\underset{\sim}{v}_c$, for $c = 1,\ldots,C$, to be a C-dimensional vector where the c-th element contains a 1 and the remaining elements contain $-1/(C-1)$.
If an unknown pixel l is found to belong to class ω_c, then y_l is coded as $\underset{\sim}{v}_c$.
Lee *et al.* (2004) develop a minimisation problem that finds the optimal set of separating functions $\underset{\sim}{f}(\underset{\sim}{x}) = (f_1(\underset{\sim}{x}),\ldots,f_C(\underset{\sim}{x}))$ subject to $\sum_{k=1}^{C} f_k(\underset{\sim}{x}) = 0$, where the solution is shown to be,

$$f_c(\underset{\sim}{x}) = \begin{cases} 1 & \text{if } c = \underset{k=1,\ldots,C}{\arg\max}\, p_{k|x}(\underset{\sim}{x}) \\ -\dfrac{1}{(C-1)} & \text{otherwise ,} \end{cases}$$

where $p_{k|x}(\underset{\sim}{x}) = \Pr(y = k \mid \underset{\sim}{x})$ is the probability that the pixel with reflectance vector $\underset{\sim}{x}$ belongs to class k. For a continuous $\underset{\sim}{x}$ space, this results in a separating function for the c-th class that takes the value 1 for the sub-domain in which the conditional probability for the c-th class is greatest, and $-1/(C-1)$ elsewhere. The authors, however, appear to assume these conditional probabilities *a priori*.

D. METHODS PROPOSED IN MACHINE LEARNING LITERATURE FOR STATISTICALLY QUANTIFYING THE ACCURACY OF SUPPORT VECTOR MACHINE PREDICTIONS

Lin, Lin and Weng (2007) proposed an implementation of binary SVM classifiers that gives class membership probabilities rather than just class labels to unknown data points via calculations that are equivalent to fitting a logistic regression model to the estimated discriminant function values. Wu, Lin and Weng (2004) extend the class probabilities to the multi-class case.

Zhang *et al.* (2012) show that the multi-class support vector machines (MSVMs) proposed by Lee *et al.* (2004) can be viewed as a maximum *a posteriori* estimation procedure under a suitable probabilistic interpretation of the classifier. They assume the following conditional probability distribution,

$$p_{c|\underset{\sim}{f}_{(-c)}(\underset{\sim}{x})} = \Pr\left(y_l = c \mid \underset{\sim}{f}_{(-c)}\left(\underset{\sim}{x}_l\right)\right) \propto \exp\left\{-\sum\nolimits_{k\neq c}^{C_t}\left(f_k\left(\underset{\sim}{x}_l\right)+\frac{1}{C_t-1}\right)_+\right\}$$

where $\underset{\sim}{f}_{(-c)}\left(\underset{\sim}{x}_l\right)$ is the vector of optimal separating functions defined earlier with the c-th element, $f_c\left(\underset{\sim}{x}_l\right)$, removed and $\left(\mathcal{A}\right)_+ = \mathcal{A}$ if $\mathcal{A} > 0$, and $\left(\mathcal{A}\right)_+ = 0$ otherwise. The intuition behind this probability distribution is that, if we assume $y_l = c$ and the optimal functions correctly label the unknown pixel l to class ω_c, then $f_k\left(\underset{\sim}{x}_l\right) = -1/\left(C_t - 1\right)$ for all $k \neq c$ and thus $p_{c|\underset{\sim}{f}_{(-c)}(\underset{\sim}{x})} \approx 1$. If, however, the optimal separating functions incorrectly assign pixel l to class $\omega_{c'}$, then $f_{c'}\left(\underset{\sim}{x}_l\right) = 1$ and $f_k\left(\underset{\sim}{x}_l\right) \leq -1/\left(C_t - 1\right)$ for all $k \notin \{c, c'\}$ and $p_{c|\underset{\sim}{f}_{(-c)}(\underset{\sim}{x})} \to 0.37$ as C_t gets large. It is noted that this choice of probability function gives a relatively high probability even when the unit is misclassified.

Zhang *et al.* (2012) then develop a hierarchical Bayesian model for posterior inference and prediction by exploiting the data augmentation methodology. This is achieved through latent variable representation for the class labels. It also makes the conditional independence assumption between the actual class membership variable and the SVM classification function explicit by linking the latent variable to the SVM kernel matrix through a linear model. The evaluation presents a comparison of test error rates with other approaches in machine learning but does not, however, discuss Bayesian credible intervals nor carry out a simulation to assess the relative bias and mean squared errors (MSEs). Another disadvantage of the method is that the Metropolis-Hastings algorithm requires the calculation of the determinant of the $n \times n$ kernel matrix at each step.

Kruppa, Liu, Biau, Kohler, Koenig, Malley and Ziegler (2014) and Mentch and Hooker (2014) are very recent additions to the machine learning literature that attempt to develop a statistical inferential basis for supervised machine learning methods. We have not had a chance yet to thoroughly assess their statistical merits and applicability to our statistical estimation problem, but on first inspection, we don't believe what is proposed in these two papers meet our needs. Papers such as Kruppa *et al.* (2014) and Mentch and Hooker (2014) demonstrate that putting machine learning techniques on a solid statistical inferential foundation is an active area of research, presently.

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

Produced by the Australian Bureau of Statistics

1352.0.55.144 • RESEARCH PAPER: METHODOLOGICAL APPROACHES FOR UTILISING SATELLITE IMAGERY TO ESTIMATE OFFICIAL CROP AREA STATISTICS (METHODOLOGY ADVISORY COMMITTEE) • June 2014