Assessment of Accuracy: Systematic Reduction of Training Points for Maximum Likelihood Classification and Mixture Discriminant Analysis (Gaussian and *t*-distribution)

Michaela Ritchie[1,2] , Pravesh Debba[3,4] , Melanie Lück-Vogel[5,6], Victoria Goodall[2,7]

[1] Council for Scientific and Industrial Research, Modelling and Digital Science, Pretoria, South Africa, mbeckley@csir.co.za
[2] Nelson Mandela University, Department of Statistics, Port Elizabeth, South Africa
[3] Council for Scientific and Industrial Research, Built Environment, Spatial Planning and Systems, Pretoria, South Africa
[4] University of the Witwatersrand, School of Statistics and Actuarial Science, Johannesburg, South Africa
[5] Council for Scientific and Industrial Research, Natural Resources and the Environment, Stellenbosch, South Africa
[6] Stellenbosch University, Department of Geography and Environmental Studies, Stellenbosch, South Africa
[7] Nelson Mandela University, Zoology Department, Centre for African Conservation Ecology, Port Elizabeth, South Africa

DOI: http://dx.doi.org/10.4314/sajg.v7i2.3

Abstract

Remote sensing provides a valuable tool for monitoring land cover across large areas of land. A simple yet popular method for land cover classification is Maximum Likelihood Classification (MLC), which assumes a single normal distribution of the samples per class in the feature space. Mixture Discriminant Analysis (MDA) is a natural extension of MLC which can be used with varying distributions and multiple distributions per class, which simplifies the classification process tremendously. We compare the accuracies of MLC and MDA (using a Gaussian and t-distribution) as the number of training points are systematically reduced in order to simulate varying reference data availability conditions. The results show that the more robust t-distribution MDA performs comparatively with the Gaussian MDA and that both outperform MLC when sufficient training points are available. As the number of training points increases the MDA accuracies increase while the MLC accuracy stagnates. At very low numbers of training samples (ranging from 22 to 169 dependent on the class), there is more variability in terms of which method performs best.

1. Introduction

Remote sensing offers a cost effective manner in which to monitor large stretches of land (Khatami et al., 2016) and has found use in many fields including forest monitoring, land change classification and change detection, natural hazard assessment, agriculture, climate dynamics, urban monitoring and ship monitoring (Khatami et al., 2016; Deng & Wu, 2013; Brusch et al., 2010). Given the current challenges relating to climate change and urban expansion, this ability to perform large scale land cover monitoring is becoming increasingly important for both environmental and spatial planning (Dewan & Yamaguchi, 2009).

While many land cover classification approaches exist, one of the most popular is Maximum Likelihood Classification (MLC) (Al-Ahmadi & Hames, 2009). MLC is an older classifier, however, it remains popular due to its simplicity. MLC does however suffer from issues relating to the assumptions of unimodality and often a Gaussian distribution of the data (Campbell & Wynne, 2011). These issues hinder the accuracy and effectiveness of the method. While some studies have made use of modified MLC methods (Mather, 1984; Maselli et al., 1995), a simple alternative, which can be seen as an extension of MLC, is Mixture Discriminant Analysis (MDA) which addresses the issues of assumed unimodality and Gaussian distribution (Ju et al., 2003).

There also exists, however, a challenge of training and validation data availability as field data may be expensive and/or time-consuming to obtain (Campbell & Wynne, 2011) and this may result in a low number of reference points which may affect the validity of the study (Congalton, 1991). Thus, it is important to understand how various levels of data availability affect the accuracy of the methods to be used. While many studies have considered effect of training sample selection on accuracy (Li et al., 2014; Jin et al., 2014; Millard & Richardson, 2015), this study focuses on assessing the accuracy of both MLC and MDA (using Gaussian and t-distributions) across varying levels of data availability in order to make recommendations regarding their use under such conditions. This study aims to make recommendations regarding the thresholds for the number of training points required to make the most of the advantages that MDA has over MLC.

2. Maximum Likelihood Classification (MLC)

MLC is a parametric pixel-based classifier which assigns classes based on the probability of belonging to a class. The training data are used to generate a single distribution (Gaussian is traditional especially in remote sensing software) to represent each class. Bayes' rule is then used to generate the posterior probability of belonging to each class. The class which has the highest probability for a pixel is the class to which that pixel is assigned (Tso & Mather, 2009).

Mathematically the posterior probability for an observation y belonging to a class k can be calculated using Bayes' rule as:

$$P(K = k|\boldsymbol{y} = \boldsymbol{Y}) = \frac{P(K = k)P(\boldsymbol{Y} = \boldsymbol{y}|K = k)}{\sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}} P(K = k)P(\boldsymbol{Y} = \boldsymbol{y}|K = k)}, \qquad [1]$$

where observation y is the vector of spectral bands for the observed pixel, *P(K=k)* is the prior probability of class *k*, *P(**Y**=**y**|K=k)* is the probability that observation ***y*** belongs to class *k*. This probability is calculated from the probability density function which is fitted to each class (Tso & Mather, 2009). The prior probabilities are often taken to be non-informative, that is every pixel has an equal probability of belonging to each class.

MLC does, however, come with the assumptions that each class is unimodal and distributed according the class chosen (usually Gaussian) (Campbell & Wynne, 2011). In practice, this is not always the case as even within a single plant species you can have variation due to age and/or canopy illumination effects (shadows) (Ju et al., 2003). To meet the assumptions within MLC, the user would need to split the training data for each class into separate classes but as this is impractical, it is often ignored, violating the assumptions which leads to lower accuracies (Hogland et al., 2013). Alternative methods such as the use of multiple endmembers in Spectral-Angle-Mapper have also been used to address this variation (Cho et al., 2010).

3. Mixture Discriminant Analysis (MDA)

MDA is a classifier which can be seen as an extension of MLC (Ju et al., 2003). In MDA, the observed classes (i.e. the land cover classes) are treated as mixtures of unobserved sub-classes. The method fits a single distribution to each of these sub-classes rather than to each class as in MLC. The mixture of the sub-class distributions is known as a finite mixture model. Various distributions can be fitted and the user is not limited to the Gaussian distribution. MDA addresses the disadvantages of MLC relating to unimodal and Gaussian distributed data while maintaining the simplicity of interpretation and implementation (Ju et al., 2003). The finite mixture models used within MDA are capable of modelling arbitrarily complex distributions (Figueiredo & Jain, 2002).

Considering a more general case of MDA than the Gaussian case considered by (Hastie & Tibshirani, 1996), we suppose we have K classes and each class can be divided into G_k sub-classes. Suppose we also have training data, $\boldsymbol{X}_k$, for each class and a set of observations $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_n\}$ for which the true class is unknown and that we wish to classify. Thus the problem is to create a classifier using the training data to label the observations $\boldsymbol{Y}$.

As mentioned, MDA makes use of a finite mixture model to generate a multi-modal distribution for each class. The resulting probability of an observation $\boldsymbol{y}$ given class k is calculated as:

$$P(\boldsymbol{Y} = \boldsymbol{y} | K = k) = \sum_{j=1}^{G_k} \pi_j f_j(\boldsymbol{y}), \qquad [2]$$

where π_j is the mixing proportion of sub-class j of class k and $f_j(\boldsymbol{y})$ is the probability density function associated with subclass j of class k (MacLachlan & Peel, 2000).

As with MLC, the EM algorithm is often used to estimate the probability density functions using the training data $\boldsymbol{X}_k$, however, other options such as numerical optimisation are available (MacLachlan & Peel, 2000; Adortse, 2016). The resulting probabilities are used in conjunction with Bayes' rule (Equation [1]) to calculate the posterior probability of class membership for an observation. The class assignment works as in MLC where the pixel is assigned to the class with the highest posterior probability.

In this study, we consider both a Gaussian MDA and a *t*-distribution MDA, as the *t*-distribution is known to be more robust to outliers due to its heavier tails (Peel & McLachlan, 2000).

4. Methodology

4.1 Study Area

The study area for this project extends from the edge of Khayelitsha to Gordon's Bay, in the Municipality of Cape Town, South Africa (Figure 1). This area falls within the Cape Floristic Region – one of six plant kingdoms in the world – with high levels of biodiversity (City of Cape Town, 2012). However, much of the natural vegetation in this area has been degraded or lost due to the high human pressure in that area (City of Cape Town, 2012).

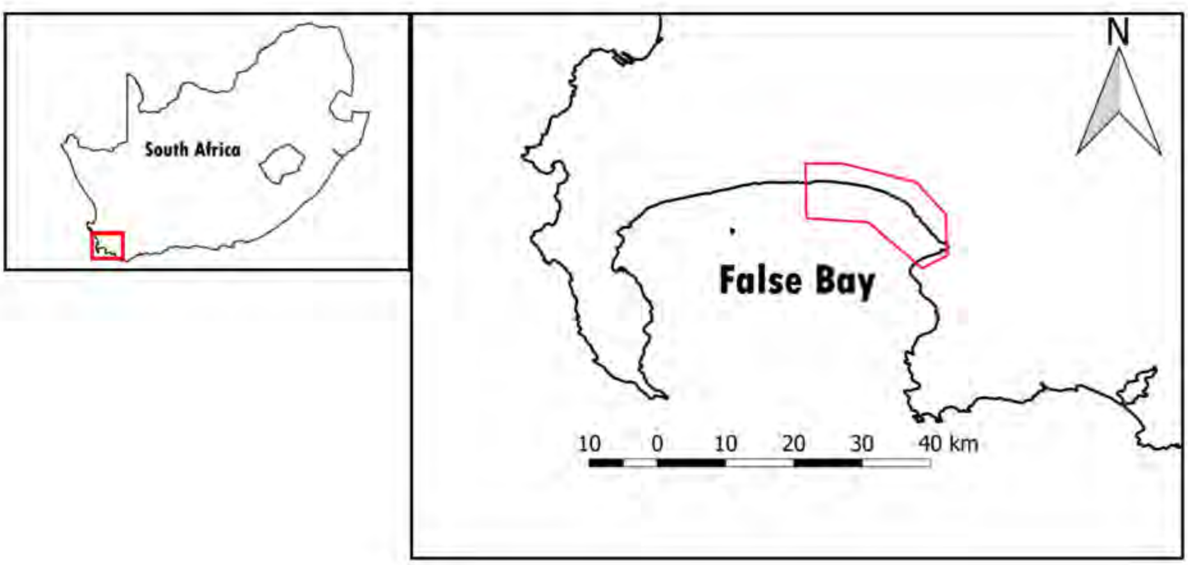


Figure 1. The study area location within South Africa and the False Bay area of Cape Town, South Africa

This area is also representative of urban coastal areas in South Africa and contains the land cover classes expected of this type of area. These classes are described in Table 1.

4.2 Data

This study makes use of Level 2a WorldView-2 satellite image acquired on the 2nd October 2014 provided by the South African National Space Agency. The image has a spatial resolution of 2m and consists of eight spectral bands, namely, coastal, blue, green, yellow, red, red edge, near infrared 1 and near infrared 2. The image was obtained in tiles and was mosaicked before atmospheric correction using ATCOR-2 software embedded in IDL (Richter & Schläpfer, 2017).

Table 1. Land Cover Class Descriptions

Class	Description
Algae	Any vegetation growing on beach rock (partly submerged or not)
Bare Ground	Any kind of uncovered soil
Built Up/Urban	Any man-made structure including but not limited to buildings, roads and bridges
Herbaceous Vegetation	Grass and other herbaceous i.e. non-woody vegetation
Shadow	Shadow caused by tall buildings and steep relief
Sparse Vegetation	Mixture of herbaceous and/or woody vegetation and bare ground and/or built up/urban and/or beach sand
Water	Any kind of open water bodies
Woody Vegetation	Trees and shrubs

4.3 Training and Validation Data

Due to the large extent of the study area, ten sub-sites were selected as core sites for selection of training and validation points (Figure 2). Pixels of each land cover type were identified in a desktop approach (digitised from the screen) for each sub-site from the WorldView-2 image based on the analyst’s prior knowledge of the study area. This approach is consistent with that of Otukei and Blaschke (Otukei & Blaschke, 2010). It is noted that not every sub-site contains all the land cover classes. The samples from each sub-site are randomly split into 70% training and 30% validation. This random split is performed 10 times to allow for 10 different runs of the methods.

To maximise the representability of the training and validation data for the classification of the whole image, the training and validation sets are composed of the previously split training and validation sets from the sub-sites. That is, the training and validation set for classifying the whole image for run one is composed of the training and validations sets for run one from each of the sub-sites. The maximum total number of training samples per class ranged from 580 to 6531. The minimum number of points per class after training point reduction was 21 (see section 4.4 below).

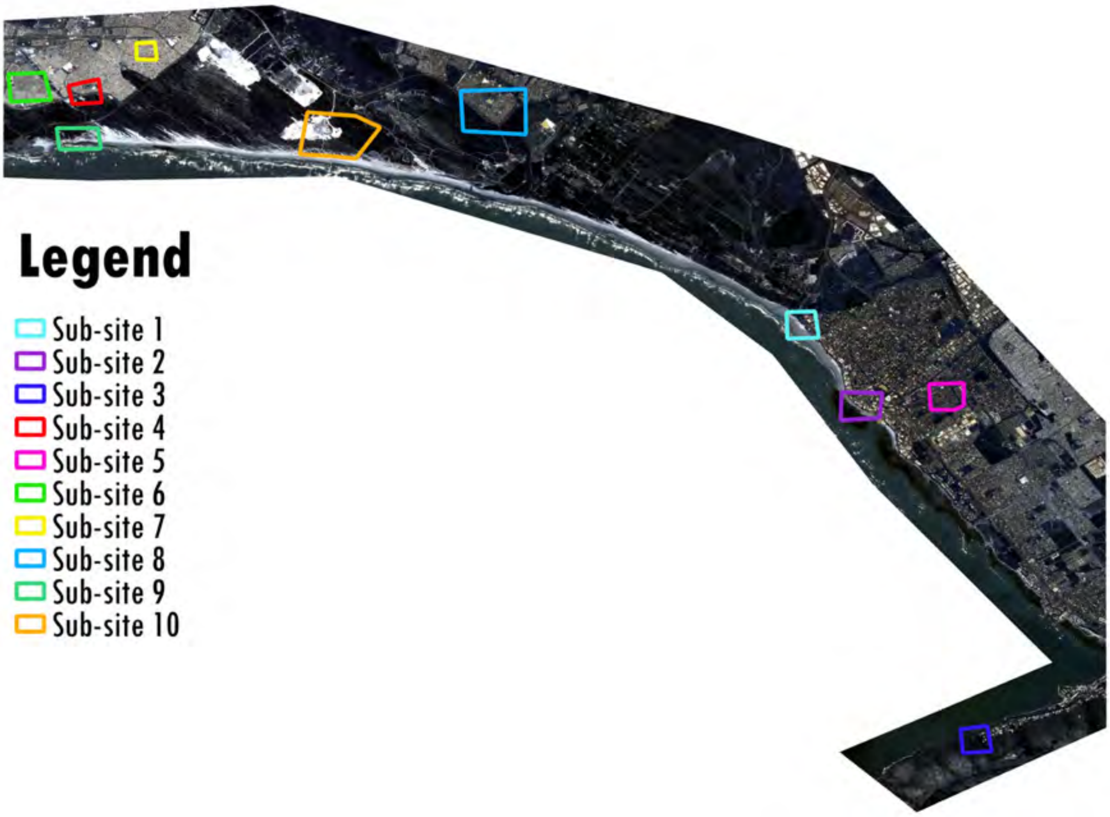


Figure 2. Sub-site locations with study area (WorldView-2 image)

4.4 Systematic Reduction of Training Points

The initial number of training points per land cover class (i.e. reduction level 0) can be seen in Table 2. This level of high training point availability is taken as the most desirable case for classifications. This high amount of training data is usually not available in real-world applications. Thus, to assess the accuracy over smaller training sets, every third point was deleted from each land cover class's training set to generate the next reduction level. The deletion of every third point was chosen as an attempt to maintain variability amongst the training points during the reduction process. Once a further reduction would result in a training set of 20 or less points, the reduction was halted for that land cover class and its training set remained the same until all classes reached that point. This is done as a minimum of 21 points is required per sub-class by the EMMIXSkew plugin (Wang et al., 2013) used to fit the distributions. This process can be seen in the pseudo-algorithm found in Algorithm 1.

The number of training points per land cover class for every reduction level can be seen in Table 2. The halting of the reduction can be seen for Algae at reduction level 8, where the training set remained the same until all classes reached their lowest allowable number of training points.

Due to the minimum of 21 points per sub-class, the maximum number of sub-classes (which we cap at ten) changes as the level of reduction is increased, this can be seen in Table 3. It is noted that reduction levels 13 and 14 only allow one sub-class per class which effectively reduces the Gaussian MDA to the standard conditions of a MLC.

Algorithm 1. Reduction in training points

```
for run in 1:maxruns{
    for class in 1:number_of_classes{
    reduction_level=0
    import training_data
    save training_data as run_class_reduction_level
    while (2/3)*length(training_data)>20 {
        reduction_level+=1
        training_data=training_data with every third row deleted
        save training_data as run_class_reduction_level
        }
    }
}
```

4.5 Expectation Maximisation (EM)-Algorithm

The EM-Algorithm is used within the EMMIXSkew plugin (Wang et al., 2013) in R (R Core Team, 2014) to fit the Gaussian and *t*-distribution MDA models. Each land cover class was fitted with a range of sub-classes from one to the number found in Table 3. In order to increase the probability of the EM algorithm converging to a global maximum rather than a local maximum or saddle point, the training was run 100 times using varying starting values. The default settings of "nrandom=10" initialisation and a maximum of 1 000 iterations per EM algorithm run were used.

Table 2. Number of training points available per class for each reduction level

Reduction Level/Class	**Built Up**	**Bare**	**Herb. Veg**	**Sparse Veg**	**Water**	**Woody Veg**	**Shadow**	**Algae**
0	6531	4592	1851	1964	3277	2277	704	580
1	4354	3061	1234	1309	2184	1518	469	370
-2	2902	2040	822	872	1456	1012	312	257
3	1934	1360	548	581	970	674	208	171
4	1289	906	365	387	646	449	138	114
5	859	604	243	258	430	299	92	76

6	572	402	162	172	286	199	61	50
7	381	268	108	114	190	132	40	33
8	254	178	72	76	126	88	26	22
9	169	118	48	50	84	58	26	22
10	112	78	32	33	56	38	26	22
11	74	52	21	22	37	25	26	22
12	49	34	21	22	24	25	26	22
13	32	22	21	22	24	25	26	22
14	21	22	21	22	24	25	26	22

Table 3: Maximum number of sub-classes per class for each reduction level

Reduction Level/Class	**Built Up**	**Bare**	**Herb. Veg**	**Sparse Veg**	**Water**	**Woody Veg**	**Shadow**	**Algae**
0	10	10	10	10	10	10	10	10
1	10	10	10	10	10	10	10	10
2	10	10	10	10	10	10	10	10
3	10	10	10	10	10	10	10	8
4	10	10	10	10	10	10	6	5
5	10	10	10	10	10	10	4	3
6	10	10	8	8	10	9	2	2
7	10	10	5	5	9	6	1	1
8	10	8	3	3	6	4	1	1
9	8	5	2	2	4	2	1	1
10	5	3	1	1	2	1	1	1
11	3	2	1	1	1	1	1	1
12	2	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1

4.6 Choosing the Number of Sub-classes

For the MLC, the model fitted was the Gaussian MDA with a single sub-class per class, as these are equivalent when the assumptions of MLC are ignored as is often done. However, for the Gaussian MDA and *t*-distribution MDA, the optimum number of sub-classes per class was chosen using the Bayesian Information Criterion (BIC) and the Integrated Likelihood Criterion (ICL). This generated four models namely Gaussian MDA with BIC (GMM BIC), Gaussian MDA with ICL (GMM ICL), *t*-distribution MDA with BIC (TMM BIC) and *t*-distribution MDA with ICL (TMM ICL). Biernacki et al., 2000 compared ICL and BIC and found that while BIC provided a sufficient estimation of the distribution, it over-estimated the number of sub-classes, while ICL provided a better estimate of the true number of sub-classes (Biernacki et al., 2000).

4.7 Classification

To convert the trained models into classifications, Python was used to assign each pixel to the land cover class for which it had the highest posterior probability. To generate the posterior probabilities, non-informative priors were used (a pixel has the same probability of being each class before any spectral information is considered). This was done for both MDA and MLC.

4.8 Validation

The validation points are taken as those remaining after the random selection of the training points (70% of all the samples). The number of validation points per class is kept constant as the number of training points is systematically reduced. The number of validation points per land cover class can be seen in Table 4.

Table 4. Number of validation points per land cover class

Class	**Number of Validation Points**
Algae	249
Bare Ground	1968
Built Up/Urban	2799
Herbaceous Vegetation	793
Shadow	301
Sparse Vegetation	844
Water	1404
Woody Vegetation	975

For assessment of the overall performance of the different classification approaches, the Kappa statistic (Campbell & Wynne, 2011), the total disagreement, quantity disagreement and allocation disagreement statistics (Pontius & Millones, 2011) were generated from the confusion matrix.

An overview of the workflow followed can be seen in Figure 3.

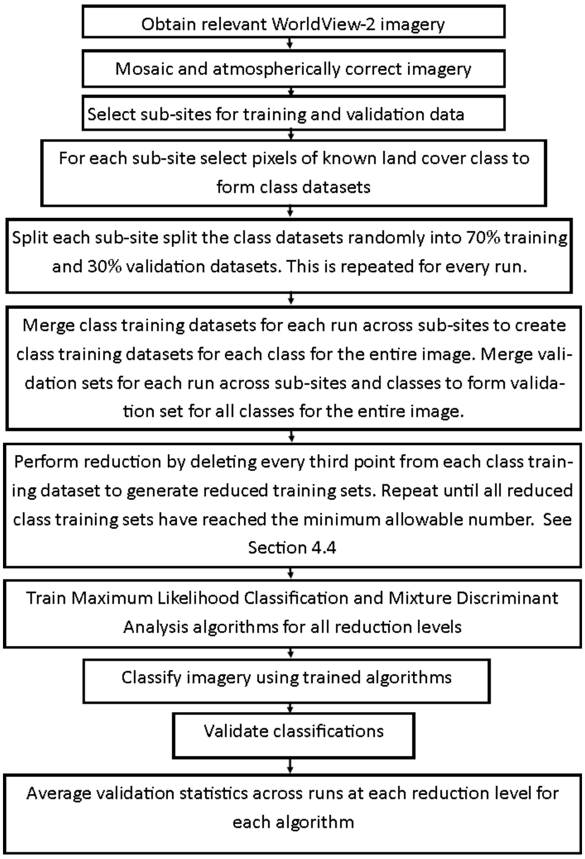

Obtain relevant WorldView-2 imagery
Mosaic and atmospherically correct imagery
Select sub-sites for training and validation data
For each sub-site select pixels of known land cover class to form class datasets
Split each sub-site split the class datasets randomly into 70% training and 30% validation datasets. This is repeated for every run.
Merge class training datasets for each run across sub-sites to create class training datasets for each class for the entire image. Merge validation sets for each run across sub-sites and classes to form validation set for all classes for the entire image.
Perform reduction by deleting every third point from each class training dataset to generate reduced training sets. Repeat until all reduced class training sets have reached the minimum allowable number. See Section 4.4
Train Maximum Likelihood Classification and Mixture Discriminant Analysis algorithms for all reduction levels
Classify imagery using trained algorithms
Validate classifications
Average validation statistics across runs at each reduction level for each algorithm

Figure 3. Workflow for generation of results

5. Results and Discussion

Each of the algorithms was run ten times, corresponding to the ten random splits of the sampled data into training and validation points. For both MDA algorithms, both BIC and ICL were used to determine the optimal number of sub-classes. For comparison of the methods, we consider the average accuracy statistics per method. These statistics are the average of the statistics across the ten runs. It is important to note that in all Figures below the number of training points decreases with increased reduction level, this means that results on the left of the graph have been trained with more points than those on the right. The number of training points available per land cover class at each reduction level can be seen in Table 2. The average Kappa statistic, average total disagreement, average quantity disagreement and average allocation disagreement can be seen in Figures 4, 5, 6 and 7 respectively.

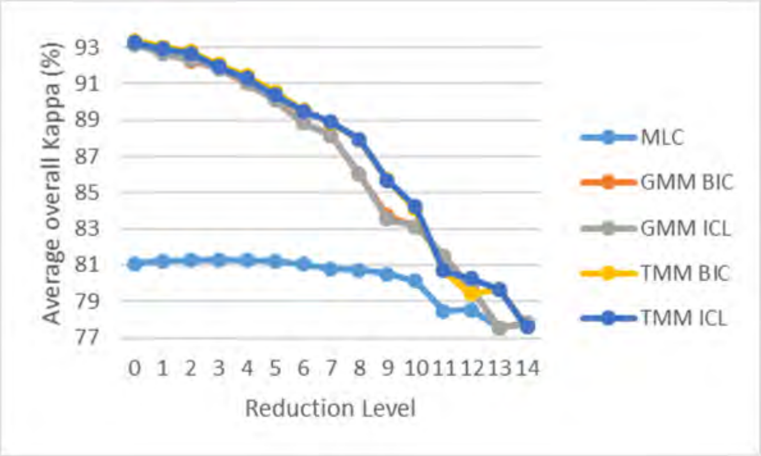


Figure 4. Average Kappa accuracy statistic per method across all training point reduction (see Table 2)

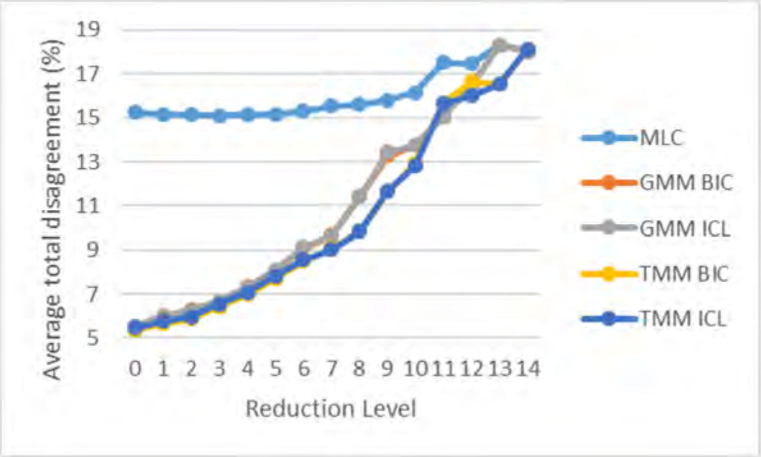


Figure 5. Average total disagreement statistic per method across all training point reduction (see Table 2)

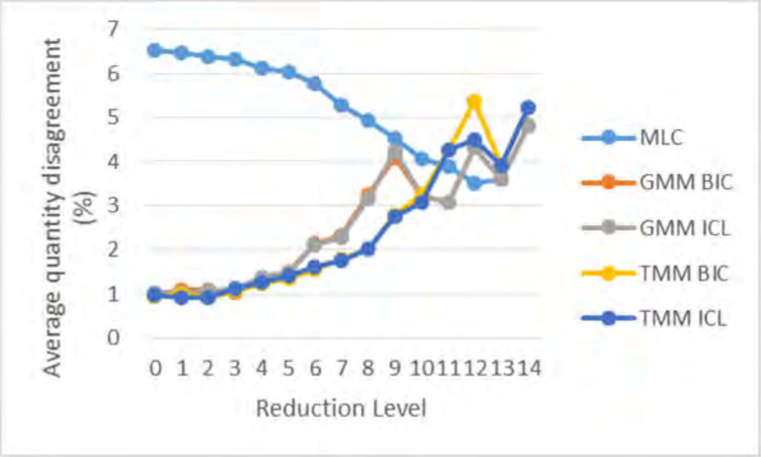


Figure 6. Average quantity disagreement statistic per method across all training point reduction (see Table 2)

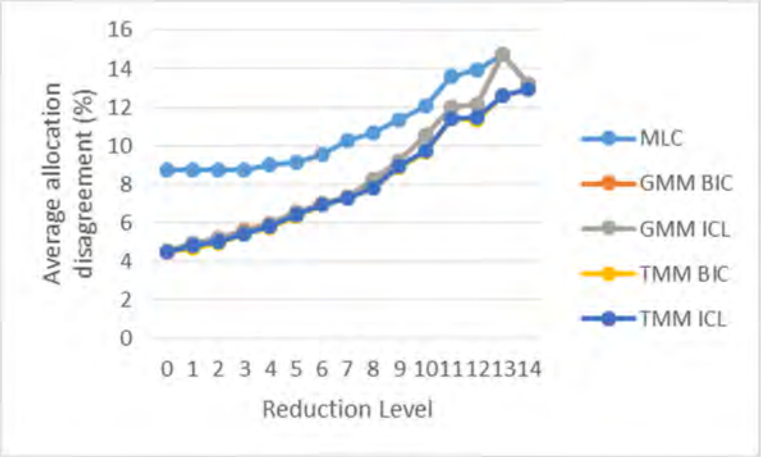
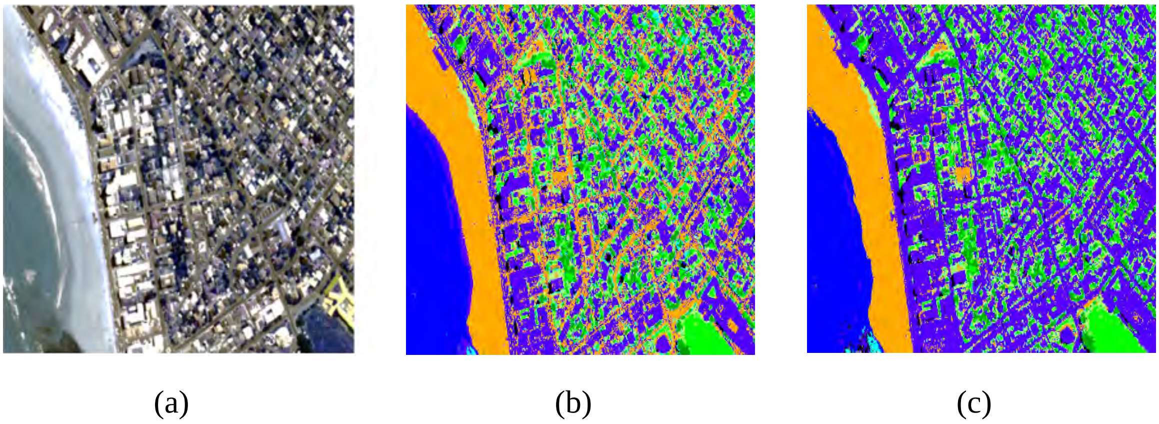
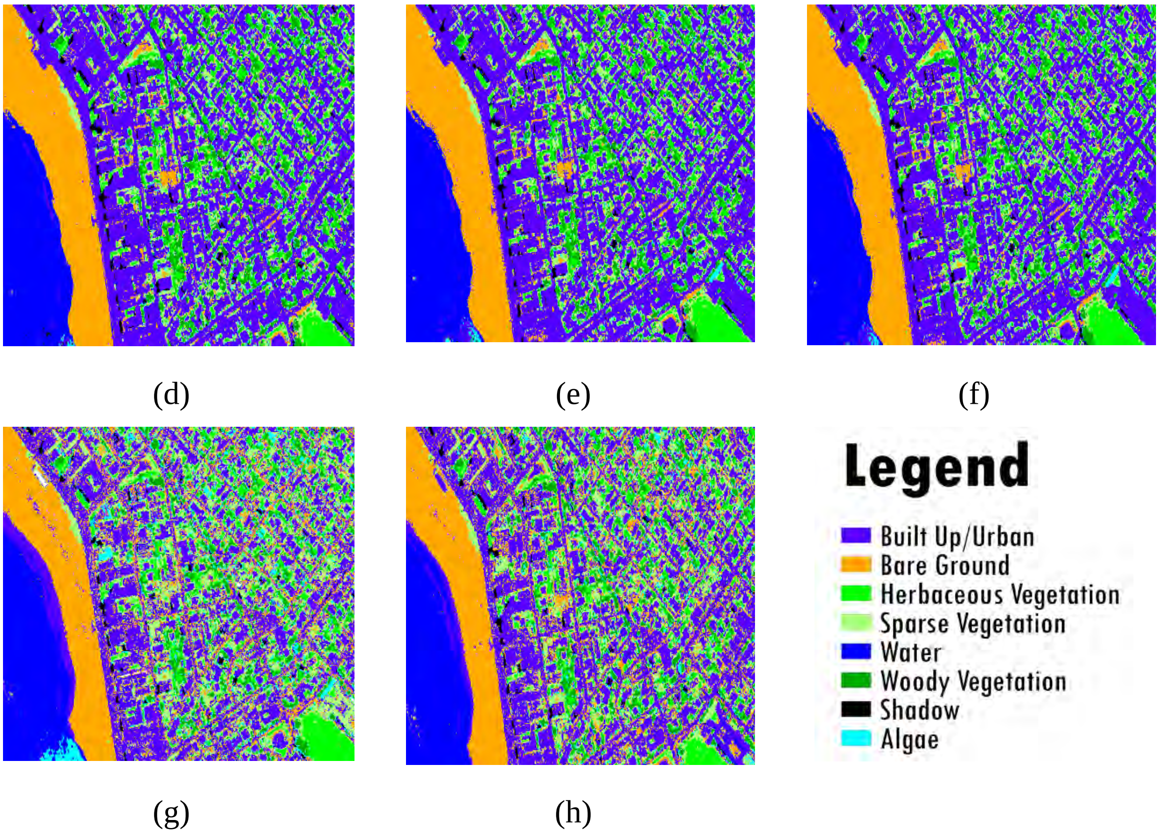


Figure 7. Average allocation disagreement statistic per method across all training point reduction (see Table 2)

When considering the Kappa statistic, the Gaussian MDA and *t*-distribution MDA methods give similar results and as reduction level increases they converge to the MLC method. The MLC method appears to have a limited capacity for accuracy and as the reduction level increases, the accuracy slowly decreases. Total disagreement shows similar trends to the Kappa statistic. From reduction level 9 there appears to be more variability in terms of the best performing method and this may be due to the small number of sample points per class, as well as the associated limiting of the maximum number of sub-classes that MDA can produce for many of the land cover classes. Majority of this variation appears to come from the quantity disagreement as the allocation disagreement remains smoother in its increase.

It is interesting to note that while the total disagreement for MLC slowly increases with reduction level, the quantity disagreement decreases and the allocation disagreement increases. That is, as the number of training points is increased, the allocation disagreement is lowered but the quantity disagreement is increased. With regards to MDA, the general trend is that the increase in total disagreement (as the number of training points is reduced) is composed of increases in both the quantity and allocation disagreements. Conversely, when the number of training points is increased, both the quantity and allocation disagreements decrease.





(d) (e) (f)

(g) (h)

Figure 8: (a) WorldView-2 image, (b) GMM MLC run one reduction level 0 classification, (c) GMM BIC run one reduction level 0 classification, (d) GMM ICL run one reduction level 0 classification, (e) TMM BIC run one reduction level 0 classification, (f) TMM ICL run one reduction level 0 classification, (g) GMM MLC, BIC and ICL run one reduction level 14 classification and (h) TMM BIC and ICL run one reduction level 14 classification for an area around sub-site two

For visual comparison, the classified images for run one at reduction levels 0 and 14 around sub-site two are shown in Figure 8 along with the WorldView-2 image for the same area. At reduction level 0, images for all five classifications are shown, while at reduction level 14, there are only two images as GMM BIC, GMM ICL and GMM MLC are identical as are TMM BIC and TMM ICL. These images demonstrate the increase confusion between classes when fewer training points (reduction level 14) are available.

6. Conclusion

To obtain the best results from MDA (both Gaussian and *t*-distribution), it is important to have sufficient training points to allow for a reasonable number of sub-classes per class. With lower numbers of training points, there is more variability between the MDA methods, however, the differences in accuracy between the MDA and MLC methods at these levels is much smaller than when more training points are available. The results of the Gaussian MDA and *t*-distribution MDA are comparable, however, the *t*-distribution MDA is known to be more robust than the Gaussian and as such, we recommend the use of *t*-distribution MDA when a reasonable number of training points is available.

The number of training points deemed to be reasonable would vary depending on the variability within the spectral signatures of a class. For a class such as built up/urban which comprises of many different surface types and colours, 150 training points may be deemed reasonable, however for a class such as herbaceous vegetation, 60 training points might be deemed reasonable. We recommend that if possible a minimum of 100 training points per class be used to allow MDA to fit at least 4 sub-classes. However, a minimum of 50 training points per class will still allow MDA an advantage over MLC as it can fit two sub-classes per class. When very few training points (less than 50) are available it is recommended to test all the methods due to variability across the runs at these levels.

In this study we ignored the unimodal Gaussian assumptions of MLC and this will have an effect on accuracy, however, this is often done in practice due to the difficulty and time taken to split the classes. In contrast, MDA performs the splitting of the classes within the algorithm itself making it easy to fit and saving on analyst time (versus splitting of classes for MLC) making it more user friendly than MLC. While this splitting of classes within the algorithm does make training slower than MLC, the time taken is not from the analyst but rather computational time.

7. References

Adortse, R., 2016. *Fitting Finite Mixture Model (FMM) to Frequency Data.* Kwame Nkrumah University of Science and Technology.

Al-Ahmadi, F. S. & Hames, A. S., 2009. Comparison of Four Classification Methods to Extract Land Use and Land Cover from Raw Satellite Images for Some Remote Arid Areas, Kingdom of Saudi Arabia. *Journal of King Abdulaziz University Earth Sciences* , Volume 20, pp. 167-191.

Biernacki, C., Celeux, G. & Govaert, G., 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern,* Volume 22, p. 719–725.

Brusch, S. et al., 2010. Ship surveillance with TerraSAR-X. *IEEE Transactions on Geoscience and Remote,* Volume 49, pp. 1092-1103.

Campbell, J. & Wynne, R., 2011. *Introduction to Remote Sensing.* The Guilford Press.

Cho, M. et al., 2010. Improving discrimination of savanna tree species through a multiple endmember spectral-angle-mapper (SAM) approach: canopy level analysis. *IEEE Transactions on Geoscience and Remote Sensing.*

City of Cape Town, 2012. *State of the environment report,* City of Cape Town.

Congalton, R. G., 1991. A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. *Remote Sensing of Environment,* Volume 37, pp. 35-46.

Deng, C. & Wu, C., 2013. “A spatially adaptive spectral mixture analysis for mapping subpixel urban impervious surface distribution. *Remote Sensing of Environment,* Volume 133, pp. 62-70.

Dewan, A. M. & Yamaguchi, Y., 2009. Land use and land cover change in Greater Dhaka, Bangladesh: Using remote sensing to promote sustainable urbanization. *Applied Geography,* Volume 29, pp. 390-401.

Figueiredo, M. & Jain, A., 2002. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Volume 24, p. 381–396.

Hastie, T. & Tibshirani, R., 1996. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society: Series B,* Volume 58, p. 155–176.

Hogland, J., Billor, N. & Anderson, N., 2013. Comparison of standard maximum likelihood classification and polytomous logistic regression used in remote sensing. *European Journal of Remote Sensing,* Volume 46, p. 623–640.

Jin, H., Stehman, S. V. & Mountrakis, G., 2014. Assessing the impact of training sample selection on accuracy of an urban classification: a case study in Denver, Colorado. *International Journal of Remote Sensing,* 25(6), pp. 2067-2081.

Ju, J., Kolaczyk, E. & Gopal, S., 2003. Gaussian mixture discriminant analysis and sub-pixel. *Remote Sensing of Environment,* Volume 84, p. 550–560.

Khatami, R., Mountrakis & Stehman, S., 2016. A meta-anlysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practioners and future research. *Remote Sensing of Environment,* Volume 177, pp. 89-100.

Li, C. et al., 2014. Comparison of Classification Algorithms and Training Sample Sizes in Urban Land Classification with Landsat Thematic Mapper Imagery. *Remote sensing,* Volume 6, pp. 964-983.

MacLachlan, G. & Peel, D., 2000. *Finite Mixture Models.* John Wiley and Sons.

Maselli, F., Conese, C., De Filippis, T. & Romani, M., 1995. Integration of ancillary data into a maximum-likelihood classifier with nonparametric priors. *ISPRS Journal of Photogrammetry and Remote Sensing,* 50(2), pp. 2-11.

Mather, P. M., 1984. A computationally-efficient maximum-likelihood classifier employing prior probabilities for remotely-sensed data. *International Journal of Remote Sensing,* pp. 369-376.

Millard, K. & Richardson, M., 2015. On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping. *Remote sensing,* Volume 7, pp. 8489-8515.

Otukei, J. R. & Blaschke, T., 2010. Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. *International Journal of Applied Earth Observation and Geoinformation,* Volume 12S, pp. S27-S31.

Peel, D. & McLachlan, G., 2000. Robust mixture modelling using the t distribution. *Statistics and Computing,* Volume 10, p. 339–348.

Pontius, J. R. & Millones, M., 2011. Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing,* Volume 32, pp. 4407-4429.

R Core Team, 2014. *R: A Language and Environment for Statistical Computing Vienna, Austria.* [Online] Available at: http://www.R-project

Richter, R. & Schläpfer, D., 2017. *Atmospheric/Topographic correction for satellite imagery: ATCOR2/3 user guide version 9.1.2.* [Online] Available at: http://www.rese.ch DLR Report Number: DLR-IB 565-01/2017 [Accessed 2017].

Tso, B. & Mather, P., 2009. *Classification Methods for Remotely Sensed Data.* CRC Press.

Wang, K., Ng, A. & McLachlan, G., 2013. *EMMIXskew: The EM Algorithm and Skew Mixture Distribution.* [Online] Available at: http://CRAN.R-project.org/package=EMMIXskew