Image from NASA: https://commons.wikimedia.org/wiki/File:Earth_Western_Hemisphere_transparent_background.png#filelinks

Remote Sensing

Classification

A common analytical process in remote sensing and image analysis is to predict landscape categories from remotely sensed data, or to derive thematic data from imagery. This can be accomplished using unsupervised or supervised classification methods to produce a variety of datasets, such as general land cover maps, forest types, vegetation community types, or wetland classifications. These methods can also be used to detect features of interest in images, such as labelling all buildings, vehicles, or individual trees. This is an evolving area of study and application, with recent advancements in computer vision, machine learning, and deep learning expanding our ability to extract such information from imagery.

The goal of this module is to provide an introduction to image classification using unsupervised and supervised classification.

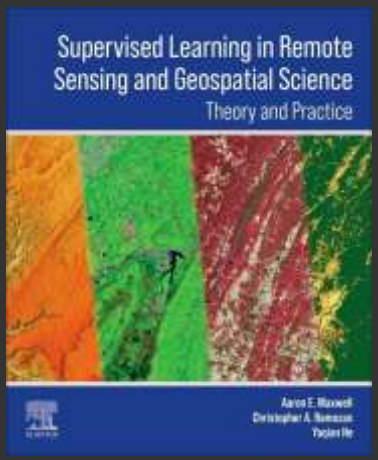
Module Topics

1. Unsupervised classification with *k*-means or ISODATA
2. Training data collection
3. Traditional supervised classification methods: minimum-distance-to-means, parallelepiped, and maximum likelihood
4. Machine learning supervised classification methods: *k*-nearest neighbors, random forest, support vector machines
5. Means to improve classification performance

2

These are the topics covered in this module.

Book Recommendation



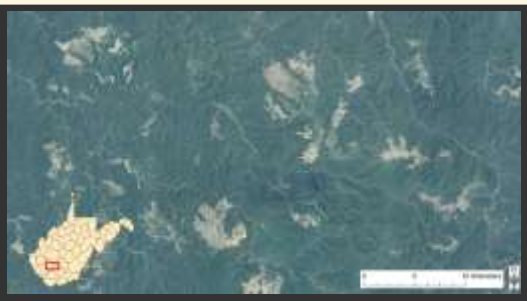
https://shop.elsevier.com/books/supervised-learning-in-remote-sensing-and-geospatial-science/e-maxwell/978-0-443-29306-1

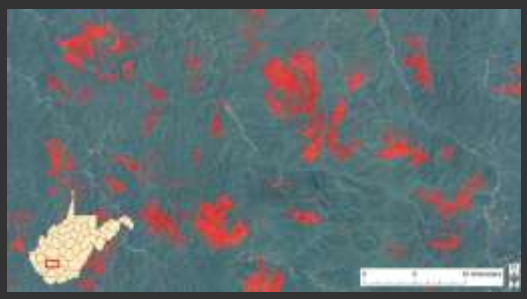
3

This book is focused on supervised learning applications in geospatial science and remote sensing.

Introduction

Image Classification





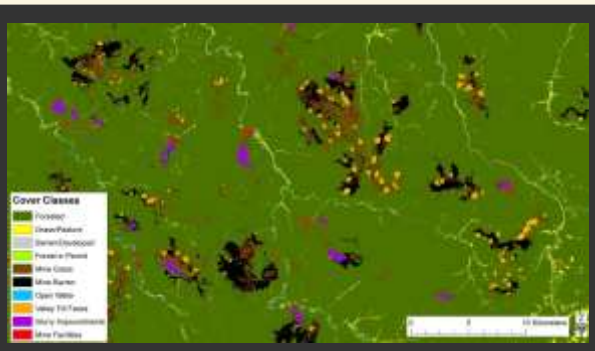


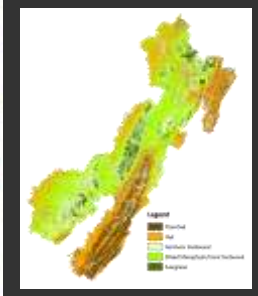
Image Bands as Predictor Variables → Thematic Classes

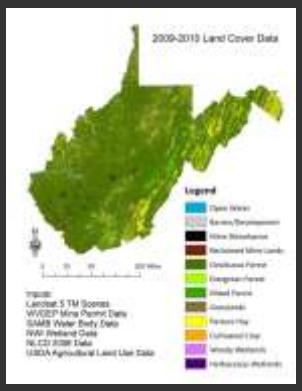
5

The images on this slide represent thematic information derived from aerial imagery, including surface mine disturbance extents and mining-related thematic classes. Generally, the goal here is to generate valuable information and derivatives from remotely sensed data.

Classification Tasks

- Land cover classification
- Forest type classification
- Vegetation classification





6

A variety of landscape classification problems exist including general land cover classification, forest type mapping, vegetation classification, and wetland mapping. This slide shows examples of some different classification products derived from remotely sensed data at varying levels of thematic detail and spatial resolution. Accurate, updateable, and consistent thematic maps are some of the most desired derivatives of remotely sensed data; as a result, this is still an active area of research in the discipline.

National Land Cover Database (NLCD)

http://www.mrlc.gov/

7

Thematic mapping from remotely sensed data has been operationalized to some degree. For example, the United States moderate spatial resolution Landsat data, along with ancillary datasets, are used to generated the National Land Cover Database on a regular basis.

National Land Cover Database (NLCD)

- 1992, 2001, 2006, 2011, 2016, 2019
- Land cover, land cover change, impervious surface, canopy cover
- http://www.mrlc.gov/

8

Different versions of the National Land Cover Database (NLCD) have been produced including data from 1992, 2001, 2006, 2011, and 2016. As part of the 2019 effort, earlier dates were filled in including 2013, 2008, and 2004.

A variety of broad land cover types are differentiated using consistent methodologies that rely predominantly on Landsat imagery.

Since these data are produced using the Landsat sensor, the spatial resolution is 30 by 30 meters.

Other than just land cover, percent impervious surface and canopy cover products are also generated.

High Spatial Resolution Land Cover

Chesapeake Bay Land Cover Data Project
http://chesapeakeconservancy.org/

https://wvgis.wvu.edu/

9

More recently, there has been an interest in developing higher spatial resolution land cover products. For example, the Chesapeake Conservancy and partners generated a 1 m spatial resolution land cover for the entire drainage basin of the bay. This large undertaking involved many partners.

In West Virginia, we have generated land cover products from the 2011 and 2016 National Agriculture Imagery Program (NAIP) orthophotography and other ancillary data sources.

High spatial resolution land cover datasets are still rare. For example, a consistent, high spatial resolution land cover of the entire United States has never been produced. This can partially be attributed to the complexity of performing these classification tasks, lack of data, cost, and computational demands.

In short, high spatial resolution land cover mapping is still an active area of research. However, recent advancements in deep learning have pushed this research area forward.

Dynamic World


scientific data

OPEN

DATA DESCRIPTOR

Dynamic World, Near real-time global 10 m land use land cover mapping

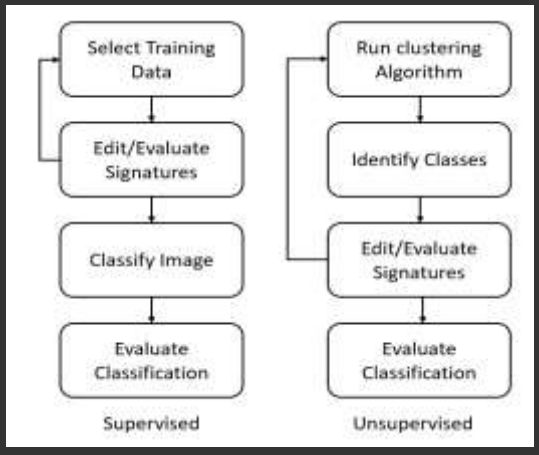
Brown, C.F., Brumby, S.P., Guzder-Williams, B., Birch, T., Hyde, S.B., Mazzariello, J., Czerwinski, W., Pasquarella, V.J., Haertel, R., Ilyushchenko, S. and Schwehr, K., 2022. Dynamic World, Near real-time global 10 m land use land cover mapping. *Scientific Data*, *9*(1), pp.1-17.

10

Deep learning methods and large datasets continue to improve our ability to generate thematic maps from multispectral imagery. For example, the Dynamic World datasets allows for land cover products to be generated from Sentinel-2 MSI imagery in near real-time. Thematic mapping technologies are progressing rapidly due to advances in big data, high performance computing, and artificial intelligence. The availability of large data archives, such as those provided by the Landsat and Sentinel-2 archives, is also key.

Supervised vs. Unsupervised

- Both require user input
- Supervised requires training data upfront
- Unsupervised requires the analyst to label the generated clusters



11

Two broad categories of image classification methods are used in remote sensing: unsupervised classification and supervised classification.

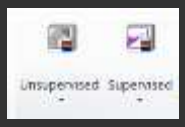
Unsupervised classification relies on data clustering techniques. An algorithm is used to cluster the data into spectral classes based on similarities between pixels. These spectral classes must then be converted to informational classes by an analyst using manual evaluation of the results.

In contrast, supervised classification requires that the analyst provide training examples upfront. These training examples are point or areal examples of the classes of interest. The examples are then combined with image data to train a learning algorithm. Once the algorithm is trained, the resulting model can be used to predict new data, such as every pixel in an image.

It is important to point out that both unsupervised and supervised classification require input from the analyst; neither are fully automated. For unsupervised classification, the user input comes at the end of the process and consists of manually assigning the spectral classes to informational classes. For supervised classification, the input comes in the form of training examples provided at the beginning of the classification process.

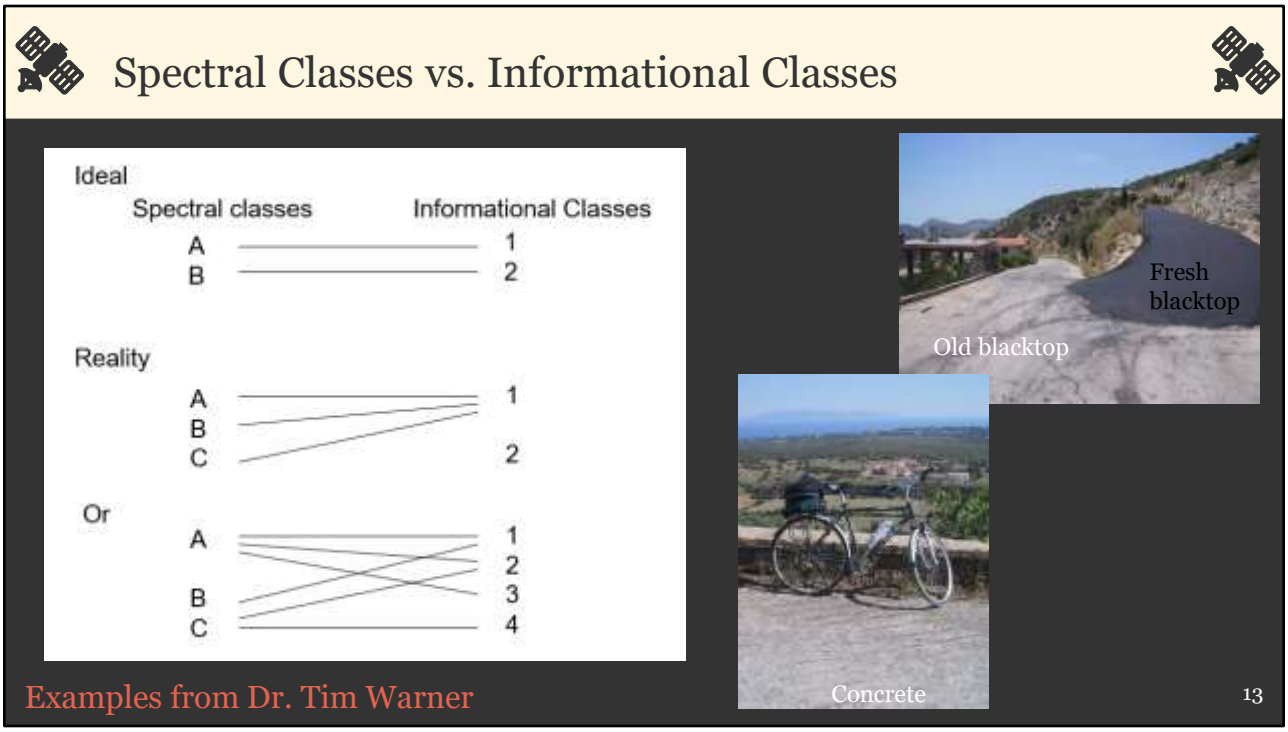
Implementations

- Erdas Imagine
- ENVI
- ArcGIS Pro
- eCognition
- R
- Python
- QGIS/Orfeo Toolbox



12

Many commercial software environments provide unsupervised and supervised classification tools including Erdas Imagine, ENVI, ArcGIS Pro, and eCognition. There are also open-source and free options, including the R, Python, and QGIS/Orfeo Toolbox data science and geospatial data science environments.



Before moving on, I want to further discuss the difference between spectral classes and informational classes.

Spectral classes represent groups of pixels with similar spectral characteristics; these are groups of pixels that have similar DN values in the available spectral bands. In contrast, informational classes represents landscape classes that have specific meaning or need to be differentiated. Forest, pastureland, cropland, developed, and water would be examples of informational land cover classes.

In the first example in the provided figure, there is a one-to-one mapping between spectral classes and informational classes. Unfortunately, spectral classes do not always align well with informational classes. In the second example, multiple spectral classes map to or aggregate to a single informational class. For example, old blacktop, which would represent a bright spectral class, and new blacktop, which would represent a dark spectral class, could combine to a single blacktop informational class. This is an issue that can be overcome by remapping the spectral classes to desired informational classes.

In the last example, some spectral classes map to multiple informational classes. This is a problem, as we cannot separate out our required

informational classes because they are too spectrally similar. An example would be concrete and old blacktop, which may have a similar spectral signature and not be well differentiated.

In short, image classification can be a complex task since all classes may not be well differentiated due to spectral similarity. This could potentially be overcome by either (1) merging classes, (2) including ancillary data, or (3) including multiple images to represent seasonality.

Unsupervised Classification

Image Bands → Spectral Classes → Informational Classes

15

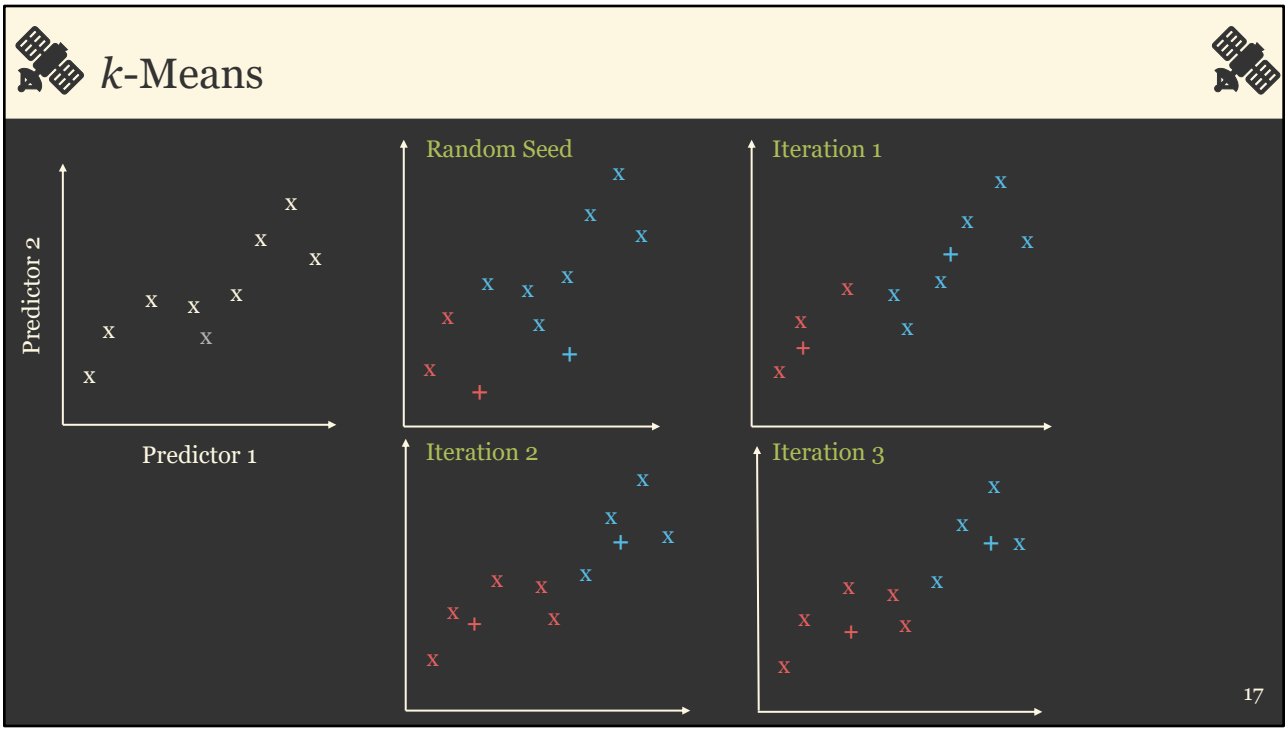
We will first discuss unsupervised classification methods. Again, the goal here is to convert an image into spectral classes by grouping pixels. These spectral classes must then be interpreted by an analyst to map them to informational classes.

k-Means

- Analyst defines the number of clusters
- Algorithm arbitrarily "seeds" or locates the cluster centers in the multidimensional measurement space
- Pixels are assigned to clusters
- Mean vectors for the clusters are determined
- Process is re-ran
- Process continues until mean vectors don't change significantly

16

One commonly used algorithm for data clustering and unsupervised classification is *k*-means. To implement this clustering method, the analyst must define the number of desired spectral classes. Generally, more spectral classes should be defined then the number of desired informational classes. The algorithm then randomly seeds or locates cluster centers in the multidimensional measurement space. Pixels are then assigned to these cluster centers based on minimal distance. This process continues through multiple iterations to move the cluster centers, or mean vectors, and reduce the overall distance between pixels and these centers. The iteration process will stop once the centers no longer change significantly. Once the final mean vectors are obtained, pixels are assigned to clusters based on which mean vector they are closest to in the multidimensional space.



This slide graphically explains the *k*-means clustering process. To begin, the graph on the left shows the location of data points in a two-dimensional feature space defined by two predictor variables. To include more predictor variables, you can map to more dimensions. For example, with three predictors, you would map a location relative to three values to a location within the volume of a cube. It is difficult to visualize more than three spatial dimensions. However, locations can be mathematically described as a multidimensional vector. This allows for clustering to be performed in many dimensions. We will stick with two dimensions here for simplicity.

First, cluster seeds are randomly placed in the feature space. Samples are then assigned to these seeds based on proximity. In this example, the red Xs represent data points that have been assigned to the red plus sign mean vector while all the blue Xs have been assigned to the blue plus sign mean vector. In order to assess how well the mean vectors are positioned, the total distance of all the data points to their assigned mean vector are summed.

In subsequent iterations, the mean vectors are moved such that the distance is minimized between each mean vector and all points that were assigned to it in the prior iteration. The data points are then reassigned to cluster centers based on proximity. This process continues until a maximum number of iterations is reached, or the mean vectors no longer need to be moved, or data points do not

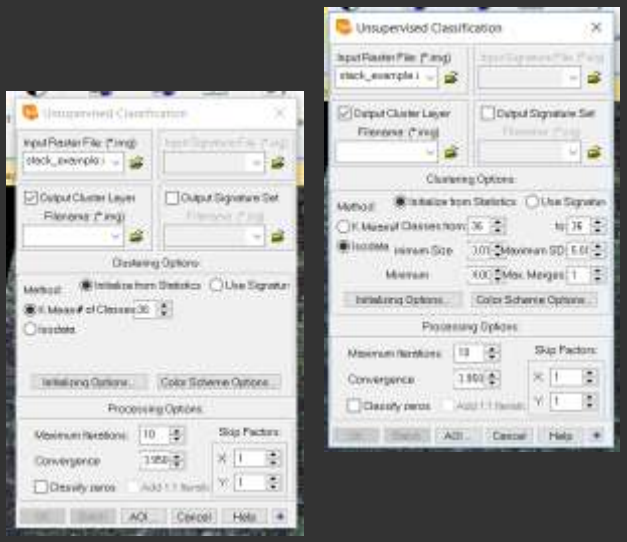
need to be reassigned to lower the sum of the distances of each data point to its associated mean vector.

There are augmentations of *k*-means in which the number of clusters or mean vectors can change as the algorithm iterates over the data. Based on a minimum number of data points that can be mapped to a cluster, a maximum within-class variance to split clusters, and a minimum pairwise distance to merge clusters, mean vectors are updated, and clusters are merged and spilt.

k-Means vs. ISODATA

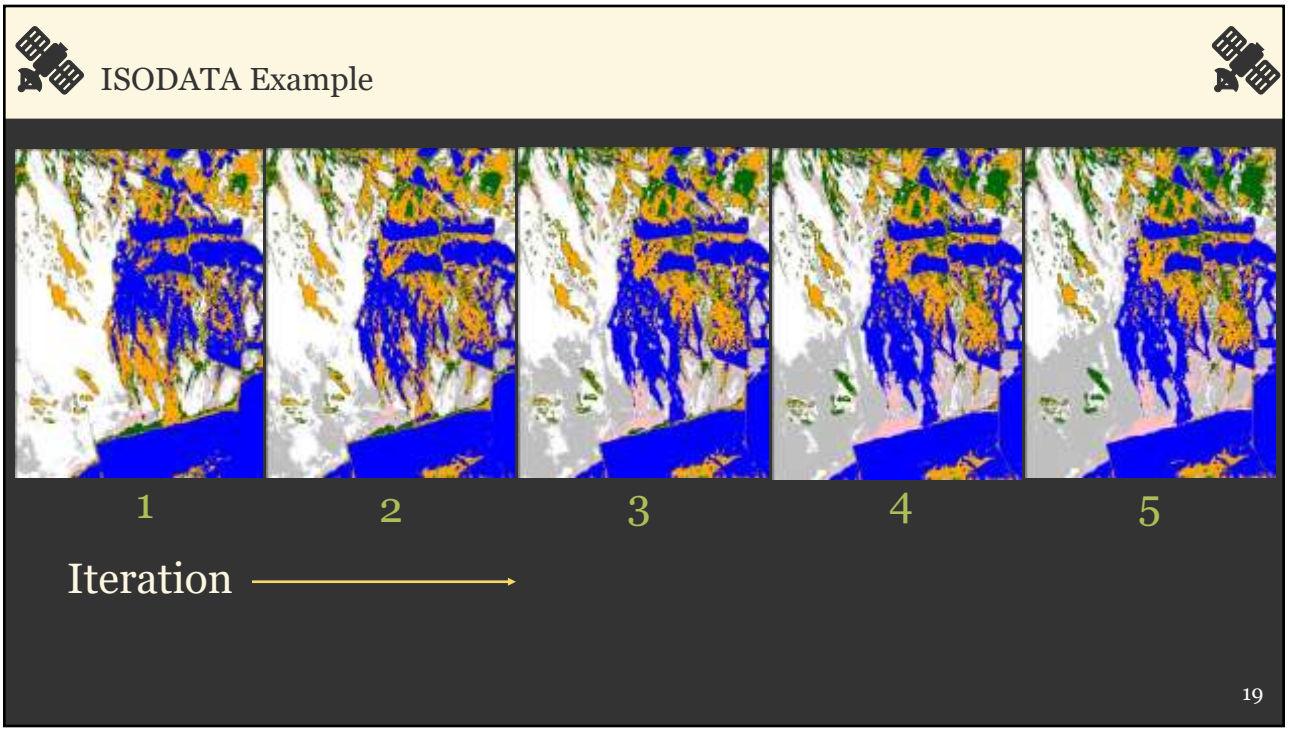
Iterative Self-Organizing Data Analysis Technique

- *k*-means: number of clusters remains the same throughout the iterations, although it may turn out later that more or fewer clusters would fit the data better.
- ISODATA: allows the number of clusters to be adjusted automatically during the iterations by merging similar clusters and splitting clusters with large standard deviations.
- Parameters
  - Desired number of clusters
  - Minimum number of samples in each cluster (for discarding clusters)
  - Maximum variance (for splitting clusters)
  - Minimum pairwise distance (for merging clusters)



18

ISODATA, or Iterative Self-Organizing Data Analysis Technique, offers an augmentation of *k*-means in which the number of cluster or mean vectors can change as the algorithm iterates over the data. Based on a minimum number of pixels that can be mapped to a classes, a maximum within-class variance to split clusters, and a minimum pairwise distance to merge clusters, mean vectors are updated, and clusters are merged and spilt.

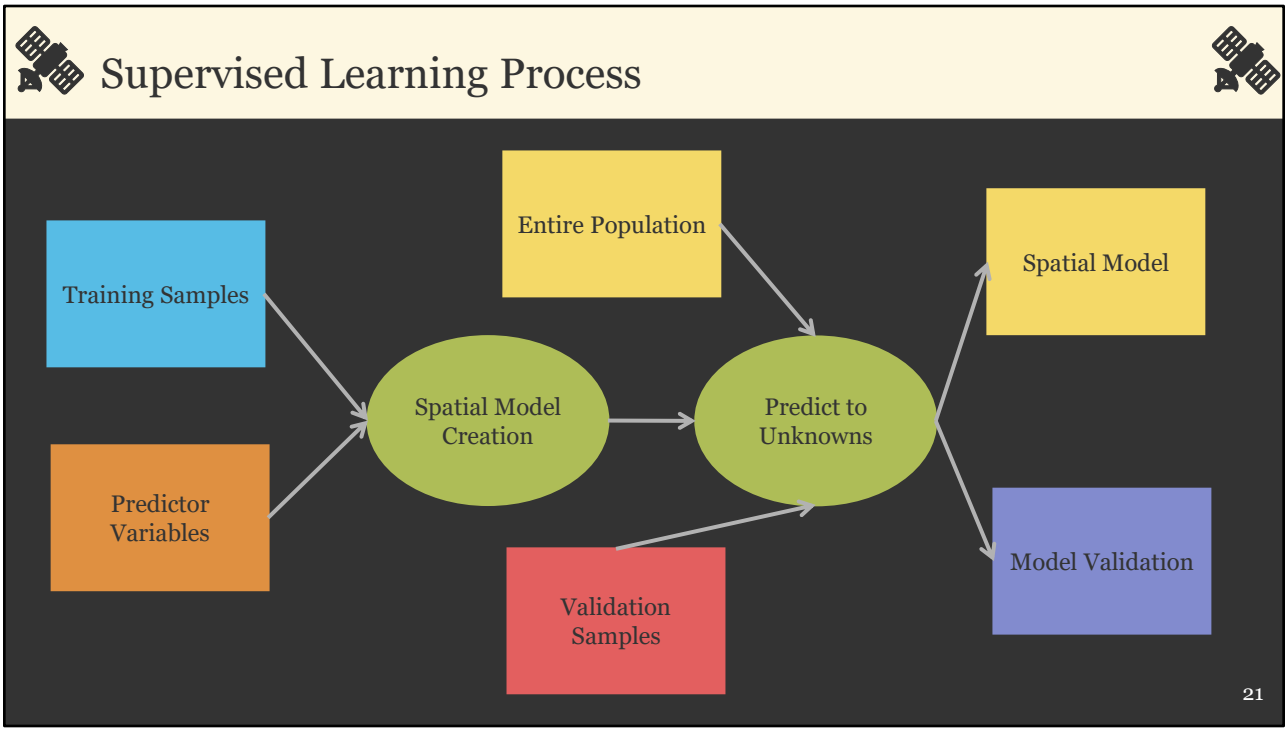


This slide shows changes in cluster or spectral class assignments throughout 5 iterations of ISODATA clustering.

Again, once pixels are assigned to spectral classes, the analyst will need to manually interpret the result to map the spectral classes to informational classes. This mapping is likely to be imperfect, with some spectral classes mapping to multiple informational

classes. There will always be some classification error present when information classes are not fully separable using the provided features or image bands.

Supervised Classification

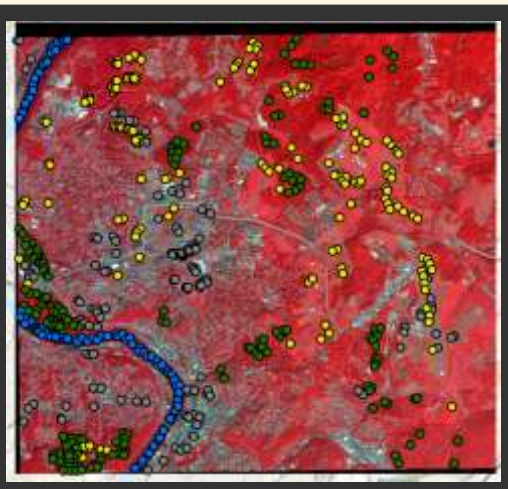


Supervised classification requires that the analyst provide training examples which will then be used by an algorithm, along with the image data, to create a model. This model is then applied to new data, or new pixels, to generate wall-to-wall classifications.

The diagram on this slide outlines the process. The algorithm will generate a model from the input training samples and predictor variables (e.g., image bands). This model can then be used to predict to the entire population (i.e., all pixels in an image) to generate a thematic map. Separate validation samples are used to assess the accuracy of the prediction. We will cover accuracy assessment in the next module, so will skip that component for now.

Training Samples

- Spatially explicit
- Based on available data, photo interpretation, and/or field work
- Representative of the classes
- Provide an adequate number of samples
- Provide and adequate number of samples per class
- Accurate



22

The training data used often has a large impact of the classification performance. Training data should be spatial explicit, meaning you know where they occur in the map space. Generally, these samples are provided as polygons, in which each pixel that intersects each polygon becomes a training sample, or individual point locations. I prefer to use point locations to minimize spatial autocorrelation in the training data. However, there is no hard rule here. Pixels that intersect training polygons can be used as opposed to points.

Training samples can be collected from existing/available data, manual photo interpretation, and/or field work. Since collecting a large number of training samples in the field using surveying methods can be expensive, time consuming, or impractical, it is common to collect training samples in the office by manual interpretation of orthophotography and ancillary datasets. It is generally a good idea to use imagery that are temporally aligned with the data being classified and that offer a higher spatial resolution. However, if high spatial resolution imagery is being classified, it is generally considered acceptable to use it as reference to collect training data.

When collecting training samples, it is important to provide samples that fully capture the spectral variability and complexity of the informational classes. For example, if you need to map blacktop, it is important to gather dark, recent

examples and brighter, older examples.

Other than just having a decent number of total samples, it is also important that you collected a variety of examples of each of the classes. The number of samples needed is not known in advance, as this will depend on the complexity of the signature, how well the class is differentiated from other classes, and the algorithm being used. Generally, I try to collect as many samples as possible given time and financial constraints.

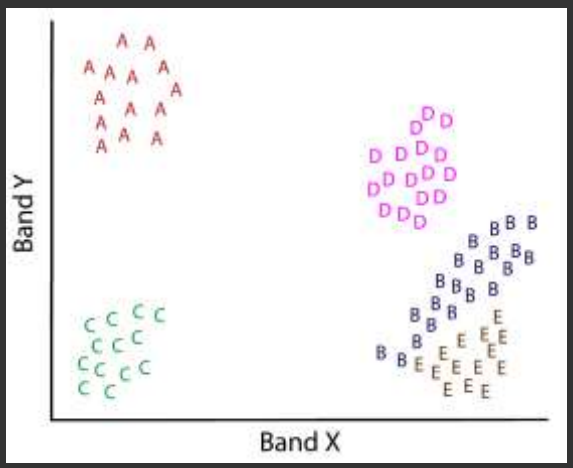
Lastly, training data should be accurately labelled and georeferenced so that the algorithm is provided a representative and accurate class signature.

To assess classification output, it is necessary to also collected unbiased and randomized validation data that do not overlap with the training examples. We will talk about this in more detail in the next module.

Once training samples have been created, pixel values and labels at these locations will be used to train a classifier.

Minimum-Distance-to-Means

- Relies on class means
- Mean or average spectral value in each band is found
- The band means comprise the mean vector
- New sample is assigned to the class that it is closest to it in the feature space based on the class mean vector
- Can specify a maximum distance to create a "unknown" class
- Does not take into account class variability
- Generally, not good if classes are close together or overlapping in the spectral space



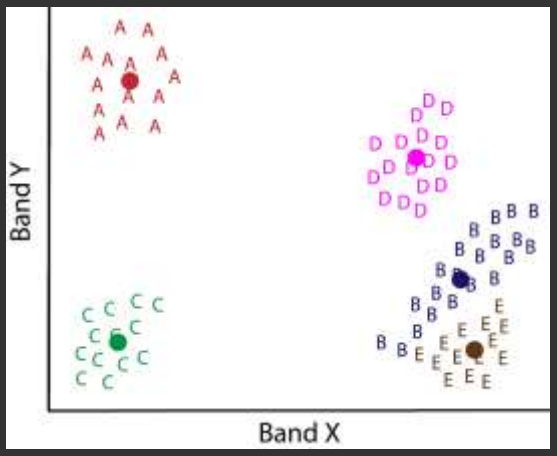
23

We will begin our discussion of supervised classification methods with traditional, parametric methods. Parametric techniques have data distribution assumptions, such as a normal distribution. They often rely on parametric summary statistics, such as mean and standard deviation.

Minimum-distance-to-means is a very simple classification algorithm. The samples in the diagram represent the training samples for five classes (A through E) plotted in a two-dimensional feature space. From these training examples, mean vectors are calculated. Progress to the next slide to see the mean vectors.

Minimum-Distance-to-Means

- Relies on class means
- Mean or average spectral value in each band is found
- The band means comprise the mean vector
- New sample is assigned to the class that it is closest to it in the feature space based on the class mean vector
- Can specify a maximum distance to create a "unknown" class
- Does not take into account class variability
- Generally, not good if classes are close together or overlapping in the spectral space

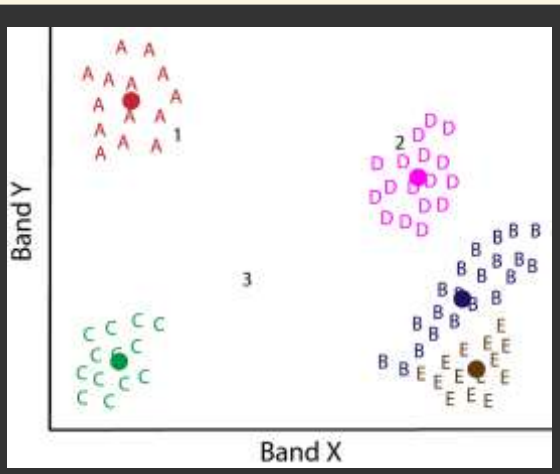


24

The mean vectors are obtained by averaging the coordinates of the training samples within each class. Here, we are showing a two-dimensional space where the class mean would be defined as the mean x and y coordinate as derived from the training samples. Again, although it is difficult to visualize, this mean vector can be calculated for many dimensions and represented mathematically.

Minimum-Distance-to-Means

- Relies on class means
- Mean or average spectral value in each band is found
- The band means comprise the mean vector
- New sample is assigned to the class that it is closest to it in the feature space based on the class mean vector
- Can specify a maximum distance to create a "unknown" class
- Does not take into account class variability
- Generally, not good if classes are close together or overlapping in the spectral space



25

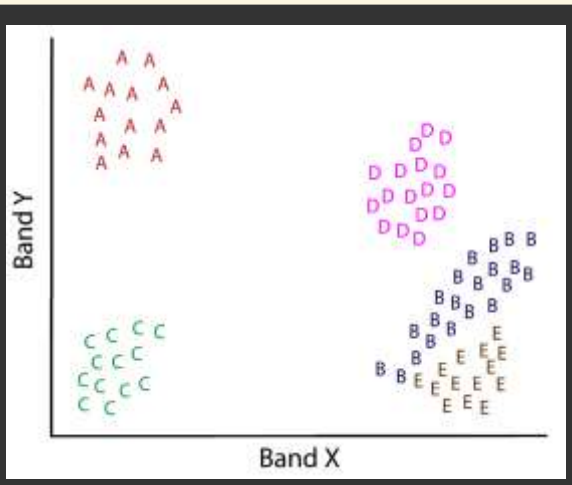
Once the mean vector for each class is determined, new pixels are then assigned to the class based on distance to the class mean vectors. Specifically, the pixel is assigned to the class to which it is closest to the class mean in the multidimensional feature space. In the example, Sample 1 would be assigned to Class A and Sample 2 would be assigned to Class D. Sample 3 is not near any class mean vector. So, it could be assigned to the nearest class, or it could be assigned as "unknown" based on a maximum distance. If a maximum distance is defined, features will be assigned to the "unknown" class if they are further than this distance from any class mean vector.

In summary, minimum-distance-to-means is based only on the central tendency of each class, as represented by the class mean vectors derived from the training samples. The variability or dispersion of the class in the spectral space is not considered.

This method is generally not a good option if classes are close together in the multi-dimensional space or have similar mean vectors.

Parallelepiped

- Relies on class range/variance
- New samples that fall within the range of a class are assigned to that class
- If a new sample is not within the data range, then it is classified as "unknown"
- If a new sample is in an overlapping area, it will be classified as "unsure" or be arbitrarily placed in one of the classes
- Correlation between bands can be an issue
- Can incorporate stepped boundaries

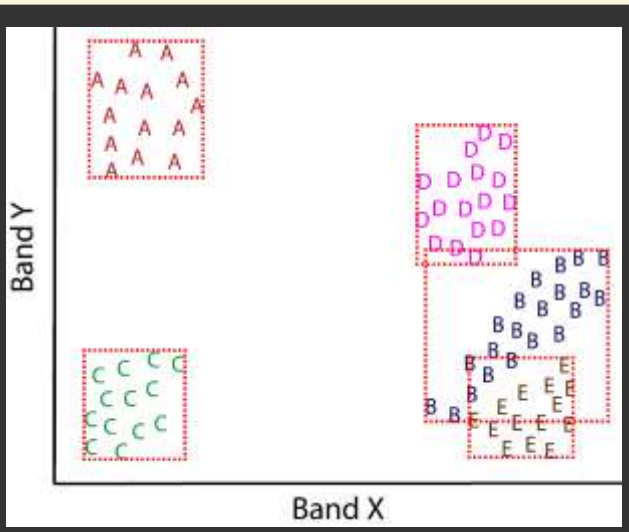


26

The parallelepiped method is based on the range of spectral values for each class as opposed to the mean spectral signatures.

Parallelepiped

- Relies on class range/variance
- New samples that fall within the range of a class are assigned to that class
- If a new sample is not within the data range, then it is classified as "unknown"
- If a new sample is in an overlapping area, it will be classified as "unsure" or be arbitrarily placed in one of the classes
- Correlation between bands can be an issue
- Can incorporate stepped boundaries



27

The boxes in the two-dimensional space represent the range of DN values for each class in two dimensions, or for two bands. In three dimensions, this would be represented as a volume. We cannot visualize this space for more than three dimensions, but it can be described mathematically.

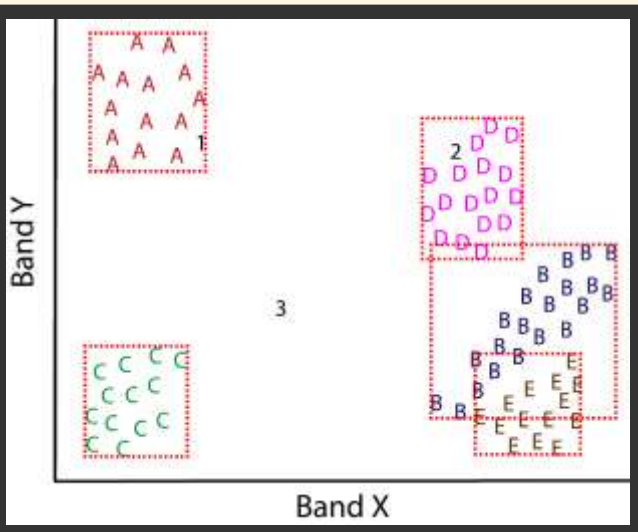
Samples that fall inside the defined class ranges will be assigned to that class. If a sample does not fall within a class range, then it will be labelled as "unknown". If class ranges overlap, such as B and E and B and D in the example, new samples can be assigned to an "unsure" class if they fall within the overlapping areas.

Note that correlation between predictor variables is an issue when using parallelepiped. For example, the B class plots as a diagonal due to correlation between Band X and Band Y. This results in a data range with a lot of empty space. So, when correlation exists between image bands, the variance or range alone is generally a poor metric for differentiating classes. Outliers can also be an issue, as they can greatly expand the class ranges.

One means to alleviate this issue to a degree is to incorporate a stair-step pattern into the ranges to remove areas in the range that contain no reference samples.

Parallelepiped

- Relies on class range/variance
- New samples that fall within the range of a class are assigned to that class
- If a new sample is not within the data range, then it is classified as "unknown"
- If a new sample is in an overlapping area, it will be classified as "unsure" or be arbitrarily placed in one of the classes
- Correlation between bands can be an issue
- Can incorporate stepped boundaries

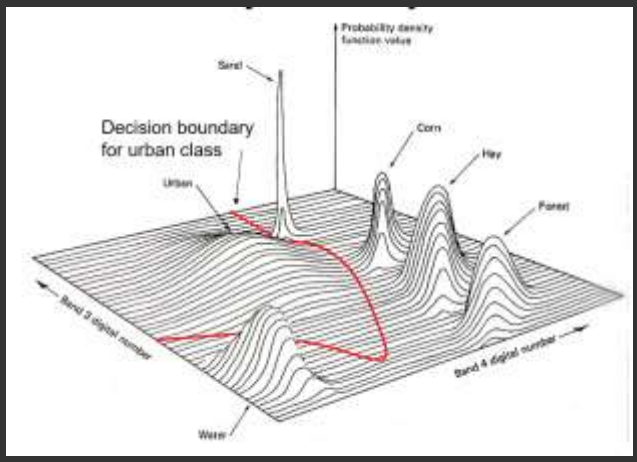


28

In this example, parallelepiped would assign Sample 1 to Class A and Sample 2 to Class D. Sample 3 would be classified as "unknown" since it does not overlap with any class ranges.

Gaussian Maximum Likelihood

- Take into account class mean and covariance
- Assumes normal distribution
- Uses mean vector and covariance matrix
- Creates a probability density matrix
- Calculates statistical probability of a given pixel being a member of a particular category
- Assigned to class of highest probability
- Can assign to an "unknown" class if all probabilities are below a threshold



Examples from Dr. Tim Warner

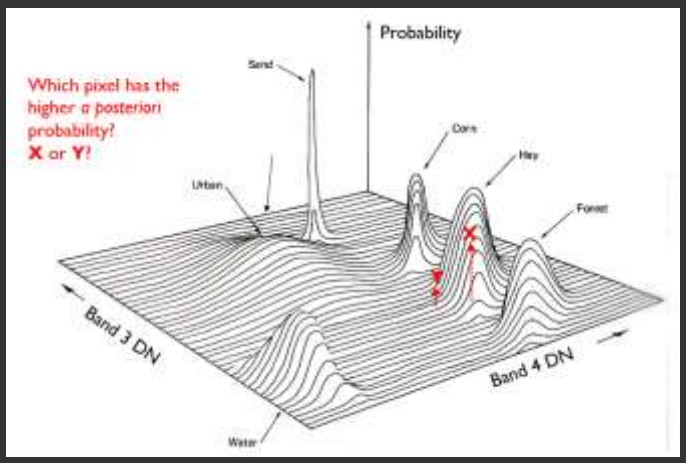
29

Gaussian Maximum Likelihood is the final parametric supervised classification technique that we will discuss. This method relies on both the class mean and variance. The class mean is defined by a mean vector, similar to minimum-distance-to-means. The variance is defined using a covariance matrix, which mathematically summarizes the covariance, spread, scatter, or relationship between two variables. Based on the parametric assumption of a Gaussian distribution for each class in each band, the probability distribution of each class can be described using only the mean vector and the covariance matrix. This allows for the probability of membership to each class to be estimated at each location in the feature space. A new sample or pixel will be assigned to the class that yields the highest probability of occurrence based on its pixel or DN values.

In contrast to minimum-distance-to-means and parallelepiped, in which a decision boundary for each class is defined in the multidimensional feature space, Gaussian maximum likelihood generates class probabilities for each class at every location in the feature space.

Gaussian Maximum Likelihood

- Take into account class mean and covariance
- Assumes normal distribution
- Uses mean vector and covariance matrix
- Creates a probability density matrix
- Calculates statistical probability of a given pixel being a member of a particular category
- Assigned to class of highest probability
- Can assign to an "unknown" class if all probabilities are below a threshold



Examples from Dr. Tim Warner

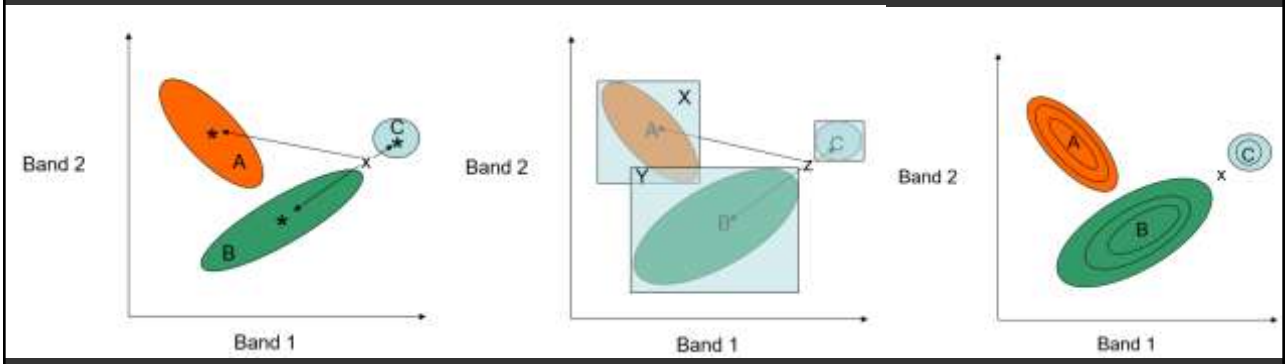
30

In the provided image, the z-dimension, or height of the surface, relates to the class probability as estimated using the Gaussian maximum likelihood method. The class mean vectors would occur at the center of the peaks and the class variability relates to the spread. For example, there appears to be less variability in the sand class than the urban class. Again, this is a two-dimensional example; however, this method can be extended to obtain class mean vectors and covariance matrices in more dimensions.

Its is also possible to assign a sample or pixel to an "unknown" class if all class probabilities are lower than a defined threshold.

Prior to the operationalization of machine learning methods, Gaussian maximum likelihood was the go-to, default classification method. However, there are some limitations. For example, the normality assumptions may not be met. If an informational class is multimodal, or has multiple probability peaks in the multi-dimensional space, then the probabilities can be misleading, and it would be more appropriate to break this class into multiple informational classes (e.g., dark blacktop and bright blacktop). These multimodal issues are also a problem for minimum-distance-to-means and parallelepiped.

Comparisons

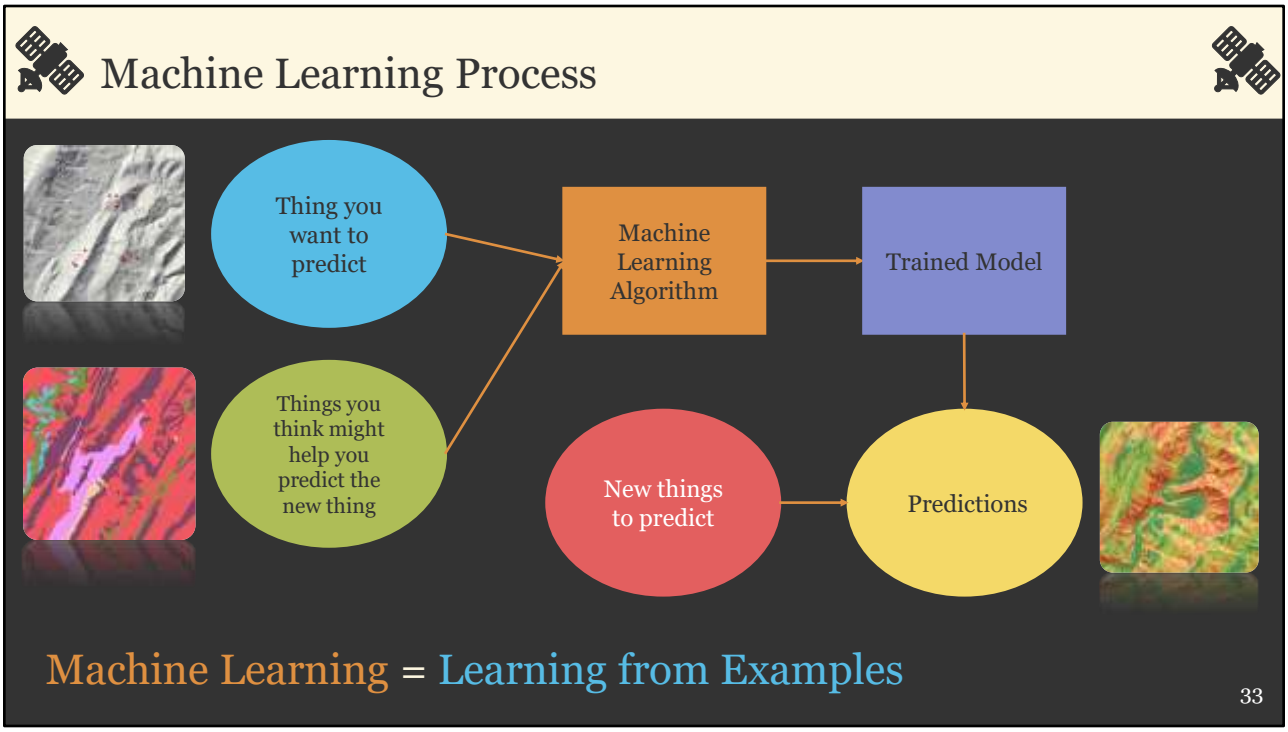


Examples from Dr. Tim Warner

31

This slide provides a comparison of the three parametric supervised classification methods discussed here. Minimum-distance-to-means only takes into account the class central tendency as represented by a mean vector. Class variability is not incorporated or considered. In contrast, parallelepiped only takes into account class variability, as represented with the class ranges. It tends to perform poorly if there are outliers in each class or correlation between bands. Prior to the operationalization of machine learning methods, Gaussian maximum likelihood was the default supervised classification method. It takes into account both class central tendency, based in mean vectors, and class variability, based on covariance matrices. This method allows for the estimate of class probabilities.

Machine Learning



Machine learning methods have now become the operational standard for image classification tasks. These methods are similar to the parametric methods that we have already discussed. However, they do not have distribution assumptions. They can generally model more complex patterns in the training data to generate more complex decision boundaries and predictions. In most circumstances, machine learning methods have been shown to outperform parametric methods.

Machine Learning

- Nonparametric
- Supervised learning
- Learn from training samples
- Generalize to unknown samples
- Generally robust
- Can model complex class signatures
- Can use a variety of predictor variables

34

This slide highlights the key characteristics of machine learning methods. These methods are nonparametric, or do not have distribution assumptions. They are often used in a supervised manor, so require training data. The models tend to be more robust to complex patterns and class signatures than parametric methods. It is also possible to incorporate a variety of predictor variables to potentially improve the classification, such as digital elevation data or ancillary variables.

Again, machine learning has matured and become the preferred method for image classification.

In the following slides, I will provide a conceptualization of three commonly used machine learning methods in remote sensing. A full discussion of machine learning is outside the scope of this course. However, an introduction is merited since these methods are now commonly used for image classification.

Machine Learning

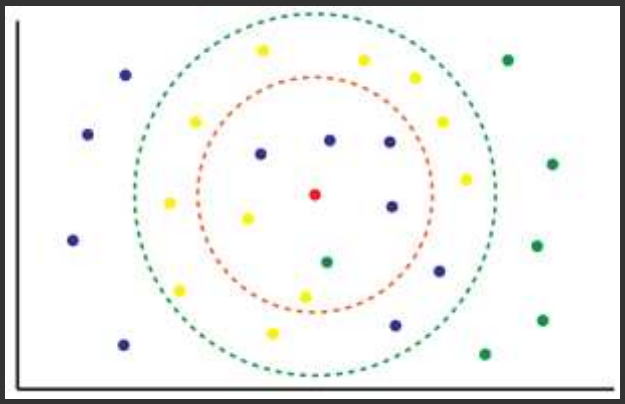
- Artificial Neural Networks (ANN)
- *k*-Nearest Neighbors (*k*NN)
- Classification and Regression Trees (CART)
- Boosted CART
- Random Forests (RF)
- Support Vector Machines (SVM)

35

This slide lists some commonly used machine learning methods. Here, I will provide a brief introduction to *k*-Nearest Neighbors (*k*-NN), random forests (RF), and support vector machines (SVM).

k-Nearest Neighbor (*k*-NN)

- Compare new sample to nearest training sample(s) in the feature space
- *k* = number of neighbors compared to
- Assign new sample to majority of the neighbors (classification)
- Use average or distance weighted average of neighbors (regression)
- Proportion of neighbors by class (probability)



36

k-NN is conceptually simple. The basic idea here is that a new observation is assigned to the class of the nearest observations in the feature space. In the provided image, the feature space is two dimensional (e.g., two image bands), but distance measures can be extended into multiple dimensions mathematically. The number of neighbors to consider is based on the *k* parameter.

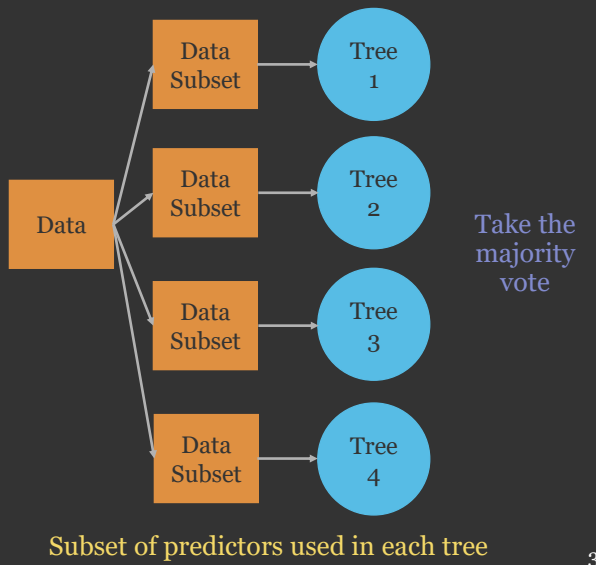
In the example, the red dot represents a new sample or pixel to be classified. The red ring represents 7 nearest neighbors. Of the 7, 4 are blue, 2 are yellow, and 1 is green. So, the new sample would get assigned to the blue class.

The green ring represents 18 nearest neighbors. From these neighbors, the yellow class is most abundant, so the new sample would get mapped to that class.

k-NN can be used to predict classes, as described here. It can also be used to predict continuous variables (regression) and class probabilities.

Random Forests (RF)

- Uses decision trees
- Uses the Gini Index of Impurity
- Ensemble decision tree method
- Uses random subset of predictor variables for splitting at each node
- Uses random subset of training data in each tree
- Attempts to reduce correlation between trees
- Ensemble of weak classifiers



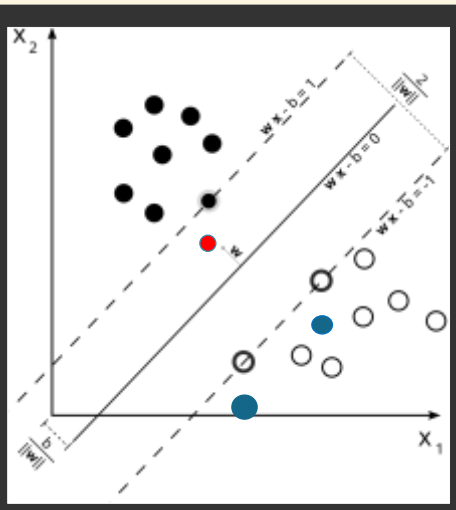
37

Random forest is based on decision trees. Decision trees use a series of binary decision rules to label samples. The decision rules are selected to homogenize the classes through recursive partitioning and are learned from the training examples.

Random forests expands this framework to include multiple decision trees that each use a subset of the training samples and input predictor variables. The goal here is to reduce the correlation between the trees and create a strong ensemble of weak classifiers. It is common to generate hundreds of trees. These trees are combined as an ensemble, and a final classification is determined based on majority voting.

Support Vector Machines (SVM)

- Works with boundary conditions
- Find the linear boundary that gives the best separation between classes (hyperplane)
- Boundary defined by support vectors, or training samples that are nearest to the boundary
- Project the data to a higher dimension
- Two class problem
  - Combine multiple classifications to separate more than three classes
- Non-separable data
  - Positive slack variable (Cortes and Vapnik (1995))
  - Cost parameter (*C*)



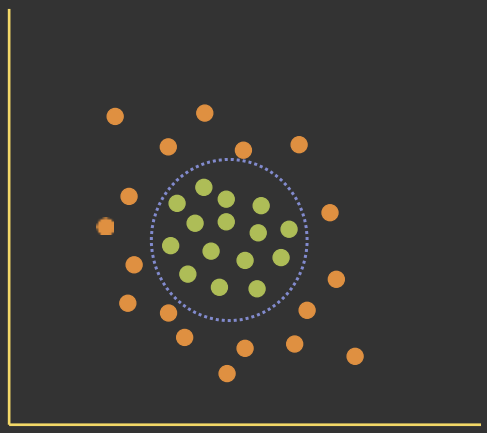
http://en.wikipedia.org/wiki/Support_vector_machine

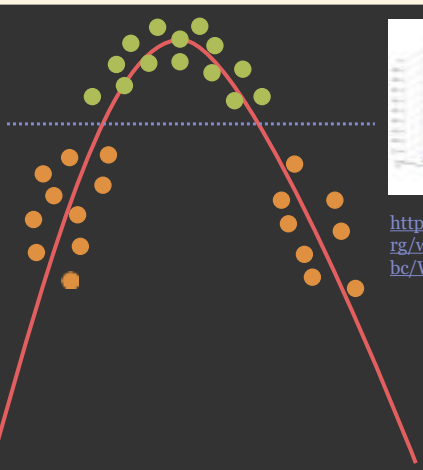
38

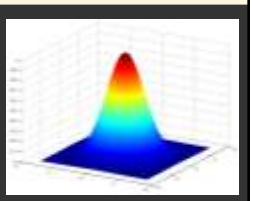
Support vector machines is a kernel-based method. The goal is to find the boundary or hyperplane that separates classes with the maximum margin. In other words, the boundary should maximize the distance between classes as defined by the training samples. Only training samples near the boundary are used to define the hyperplane and are termed support vectors.

One issue is that classes may not be linearly separable in the feature space; a line cannot be drawn in the space to separate the classes. In order to improve separability, a kernel is used to transform or map the samples to a higher dimensional space where classes are more separable.

The Kernel Trick







https://upload.wikimedia.org/wikipedia/commons/b/bc/Wiki_gauss.png

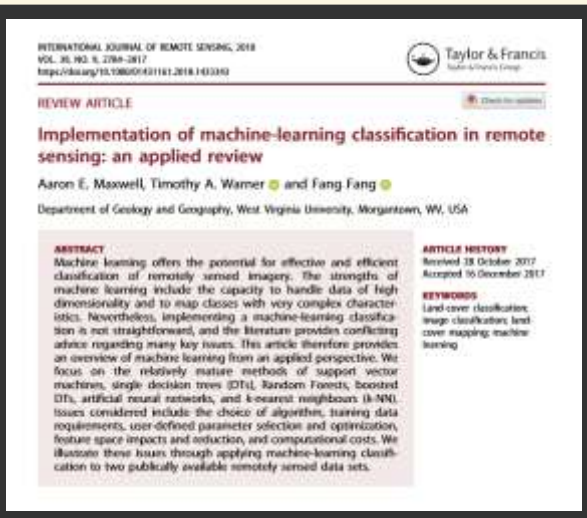
- Use the kernel trick to transform the feature space into a higher dimensional space where the features are more linearly separable

39

This slide further conceptualizes the kernel trick. Again, the goal is to project the input predictor variable space into a higher dimension in which a nonlinear boundary may become more linear and defined as a hyperplane.

As an example, imagine that the separating boundary between two classes in a two-dimensional space would best be defined as a circle, as pictured in the image on the left. The samples could be projected onto the surface of a three-dimensional curve where the separating boundary could be described as a plane. Thus, a nonlinear boundary was made more linear by projecting the data into a higher dimensional feature space.

Machine Learning in Remote Sensing



INTERNATIONAL JOURNAL OF REMOTE SENSING, 2018
VOL. 39, NO. 9, 2784–2817
https://doi.org/10.1080/01431161.2018.1433343

Taylor & Francis

REVIEW ARTICLE

Implementation of machine-learning classification in remote sensing: an applied review

Aaron E. Maxwell, Timothy A. Warner and Fang Fang

Department of Geology and Geography, West Virginia University, Morgantown, WV, USA

ABSTRACT

Machine learning offers the potential for effective and efficient classification of remotely sensed imagery. The strengths of machine learning include the capacity to handle data of high dimensionality and to map classes with very complex characteristics. Nevertheless, implementing a machine-learning classification is not straightforward, and the literature provides conflicting advice regarding many key issues. This article therefore provides an overview of machine learning from an applied perspective. We focus on the relatively mature methods of support vector machines, single decision trees (DTs), Random Forests, boosted DTs, artificial neural networks, and k-nearest neighbours (k-NN). Issues considered include the choice of algorithm, training data requirements, user-defined parameter selection and optimization, feature space impacts and reduction, and computational costs. We illustrate these issues through applying machine-learning classification to two publically available remotely sensed data sets.

ARTICLE HISTORY
Received 28 October 2017
Accepted 16 December 2017

KEYWORDS
Land-cover classification; image classification; land cover mapping; machine learning

Maxwell, A.E., Warner, T.A. and Fang, F., 2018. Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, *39*(9), pp.2784-2817.

https://www.tandfonline.com/doi/full/10.1080/01431161.2018.1433343

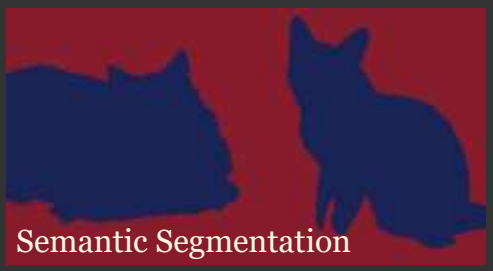
40

Again, our goal here was not to provide an in-depth discussion of supervised classification machine learning. Instead, the topic was simply introduced. Machine learning is discussed in some of our more advanced courses if you are interested.

On this slide, I have provided a link to an open-source manuscript focused on the applied application of machine learning in remote sensing.

Commonly used machine learning algorithms, such as random forests and support vector machines, have been integrated into software packages, such as Erdas Imagine, ENVI, ArcGIS Pro, QGIS/Orfeo Toolbox, R, and Python. So, these algorithms can now be readily implemented and do not require coding knowledge.

Deep Learning (DL)



41

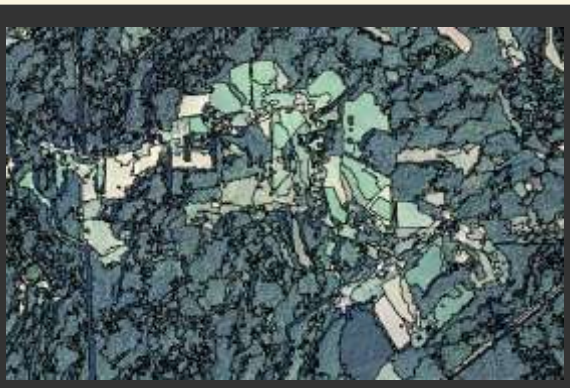
Over the last few years, research into deep learning and computer vision has resulted in many powerful algorithms for labelling images, pixel-level classification, or detection of unique instances of a class. These deep learning methods are based upon artificial neural networks. More specifically, convolutional neural networks have proven to be extremely powerful for image analysis due to their ability to learn spatial patterns in the image data.

We will introduce some deep learning and GeoAI techniques in a later module.

Additional Topics

Object-Based Classification (GEOBIA)

- Geographic Object-Based Image Analysis
- Commonly used for high spatial resolution data
- Segment image into objects or polygons
- Classify objects as opposed to pixels
- Help remove salt-and-pepper effect
- Potentially a more cartographically pleasing result
- Incorporate ancillary data, textural measures, and geometric measures



43

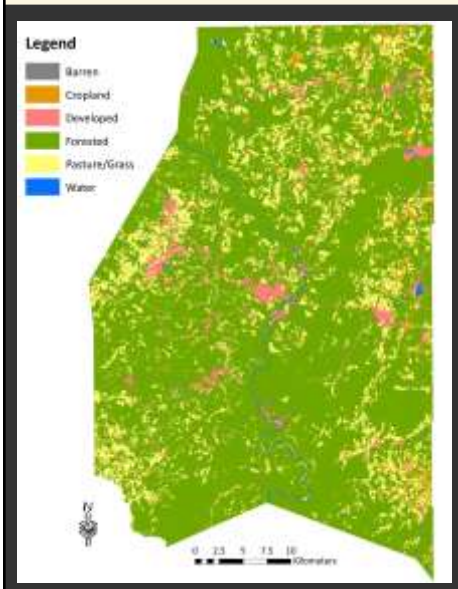
When high spatial resolution imagery are classified, it is common to aggregate pixels to polygons or image objects. This is known as geographic object-base image analysis (GEOBIA). An algorithm is used to segment the image into objects or polygons based on similarity between adjacent pixels. These objects then become the unit of analysis for subsequent classification. For each object, a variety of summary statistics can be calculated from the input imagery and ancillary data relating to central tendency, texture or variability, and object shape.

Rule sets can be used to differentiate classes based on the calculated variables. Alternatively, example, labelled objects and associated metrics can be provided as training data to train a machine learning classifier. The trained model can then be used to predict each object in the mapped extent.

GEOBIA tends to minimize the salt-and-pepper issue common in pixel-based classification, which arises when small groups of pixels are misclassified. The results tend to also be more cartographically pleasing. Although these methods are commonly used to analyze high spatial resolution data, in which the pixels are smaller than the features of interest, they have also been applied to moderate spatial resolution data, such as Landsat.

We provide a brief discussion of GEOBIA in a later module.

GEOBIA Example



Spectral	1st Order Texture	2nd Order Texture	NDVI	Geometry	LiDAR	Road Density	Structure Density
Mean Blue	SD Blue	Homogeneity	Mean	Area	Mean nDSM	Mean 1,000 m radius	Mean 1,000 m radius
Mean Green	SD Green	2nd Angular Moment	SD	Asymmetry	SD nDSM	Mean 500 m radius	Mean 500 m radius
Mean Red	SD Red	Contrast		Border Index	Minimum nDSM	Mean 250 m radius	Mean 250 m radius
Mean NIR	SD NIR	Entropy		Border Length	Maximum nDSM	SD 1,000 m radius	SD 1,000 m radius
		Dissimilarity		Compactness	Mean Slope	SD 500 m radius	SD 500 m radius
		Correlation		Roundness	SD Slope	SD 250 m radius	SD 250 m radius
		Standard Deviation		Width	Minimum Slope		
		* for all 4 bands		Elliptic Fit	Maximum Slope		
				Density	Mean Elevation		
				Rectangular Fit	SD Elevation		
				Length	Mean First Return Intensity		
				Length/Width	SD First Return Intensity		
				Volume	Homogeneity First Return Intensity		
				Shape Index	2nd Angular Moment First Return Intensity		
					Contrast First Return Intensity		
					Entropy First Return Intensity		
					Dissimilarity First Return Intensity		
					Correlation First Return Intensity		
					Standard Deviation First Return Intensity		
4	4	28	2	14	19	6	6

Total: 83 Variables

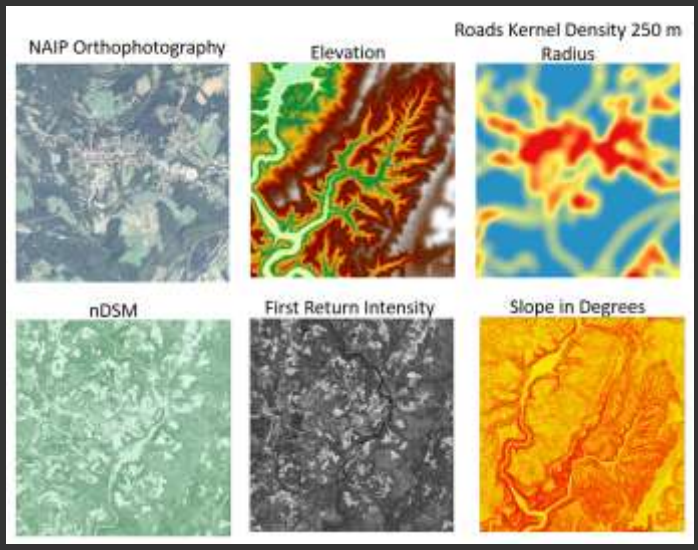
44

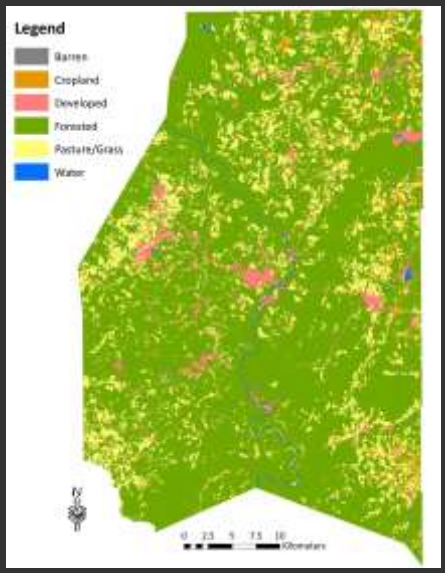
This slide shows an example classification generated using GEOBIA. The table lists the metrics or features calculated for each image object.

In order to produce the classification, a set of training objects and their associated metrics were used to train a machine learning algorithm. The trained model was then applied to classify all objects in the study area extent to produce a land cover map for the entire county.

Again, one of the benefits of GEOBIA is the ability to summarize a variety of data on each object. In this mapping project, we summarized aerial image bands to obtain measures of central tendency and texture; LiDAR data to obtain measures of canopy height, return intensity, central tendency, and texture; and the density of roads and man-made structures, which were derived from ancillary vector data.

GEOBIA Example





45

This slide shows some of the input data used in this study that were derived from imagery, LiDAR, or ancillary data.

Some Complexities

- Not all classes are spectrally separable
- Data can be suboptimal
- Training data can be difficult to collect
- It is important to consider the impact of scale
- Classes may have complex signatures
- Classes may have fuzzy definitions
- Pixels can be "mixed"
- There can be positional errors in your data

46

Image classification tasks can be very complex. This skillset can take some time to develop. It is an art and a science.

This complexity relates to many factors. First, the defined informational classes may not be spectrally separable, so image data alone may not be adequate to accurately differentiate all categories. Input data can be suboptimal for many reasons including issues of spectral, spatial, radiometric, and/or temporal resolution. Images may not be well calibrated. Inconsistencies in data are of particular concern when classification is performed on aerial image mosaics made up of images that were collected on different dates with different illuminating conditions.

Quality training data are generally very important for a successful classification. However, it is not always easy to collect a large number of quality examples that adequately characterize the classes of interest due to cost, time, and access constraints. Not all classes can be easily mapped or differentiated using manual photograph interpretation or even field work.

It is important to consider the impact of scale. Classes that are observable in a high spatial resolution image may not be observable in coarser data. It is important to think about how the data products will be used and what classes and level of spatial detail are required.

Individual classes can have complex signatures that do not honor distribution assumptions. It is also not always easy to define discrete classes. For example, the difference between “mixed developed” and “highly developed” may be difficult to quantify and explicitly define. Some map features have inherently fuzzy or gradational boundaries, such as the transition between wetlands and uplands.

For moderate to coarse spatial resolution data, individual pixels may be mixed. For example, a single Landsat 30-by-30 meter pixel could contain some forest, pastureland, and water. Thus, assigning the pixel to one class, as is commonly required to generate a thematic map, is inherently flawed.

Also, all spatial data have some degree of positional error due to issues of georeferencing. This error will impact the positional accuracy of any derived products.

Improving your results

- Re-execute with more training data
- Combine confused classes
- Incorporate ancillary data
- Use a more powerful algorithm
- Smooth or generalize results
- Make use of object-based classification
- Use a different image
- Incorporate spectral enhancement/ratios
- Incorporate measures of texture
- Use multi-temporal data
- Use multi-seasonal data
- Manually clean up results
- Consider defining a Minimal Mapping Unit (MMU)

47

If your classification results are inadequate, there are some options for potentially improving them. First, you may need to collect more training data to refine the classification model. We have found that classification output can often be improved by collecting more training data over areas that were misclassified in the prior iteration. You may also find that some classes can not be accurately differentiated and may need to be combined.

Incorporating ancillary data may help to differentiate classes that are spectrally similar. These ancillary data could include LiDAR-derivatives to characterize the terrain surface or vegetation heights. Ancillary vector data may be useful, such as man-made structures or roads. It is also possible to "burn-in" vector data into the output. It is common for final thematic maps to incorporate a combination of supervised classification results and supplemental ancillary data. For example, building footprints could be added as a "building" class.

Using a more powerful algorithm may yield better results. Unfortunately, no algorithm has been shown to be optimal for all problems. However, machine learning methods have generally been shown to consistently outperform parametric methods, like Gaussian maximum likelihood. We recommend using random forests or support vector machines as a default algorithm.

Using moving windows to smooth or generalize your output may yield more

cartographically pleasing results. Object-based classification is generally worth considering for high spatial resolution data and may yield more cartographically pleasing results than pixel-based classification by reducing the salt-and-pepper effect.

You may want to experiment with different imagery, as not all imagery are adequate for all mapping tasks. Reduced spectral resolution may limit the number of classes that can be accurately differentiated. You can also attempt to incorporate additional measures from the available imagery, such as spectral enhancements, band ratios (e.g., NDVI), or textural measures. When data are collected by sensors with a regular return interval, such as from a satellite platform, you can experiment with incorporating multiple images representing different seasons or phenology. You can also derive measures from multi-temporal data that capture seasonal patterns and variability. Such methods are of particular value when you are attempting to map features with unique seasonal changes and patterns, such as forest types or agriculture.

If you are not able to obtain acceptable results using these suggestions, you may need to manually clean up the results. Many practical, applied mapping projects still incorporate manual clean up to improve upon results obtained from supervised classification. Lastly, you should define a minimal mapping unit (MMU), which relates to the smallest features to be mapped. This can inform users as to the appropriate scale for analyzing and working with the thematic data.

Sieving

- Remove areas smaller than a certain size or area
- Good for generalizing results
- Good for mapping using a minimum mapping unit (MMU)

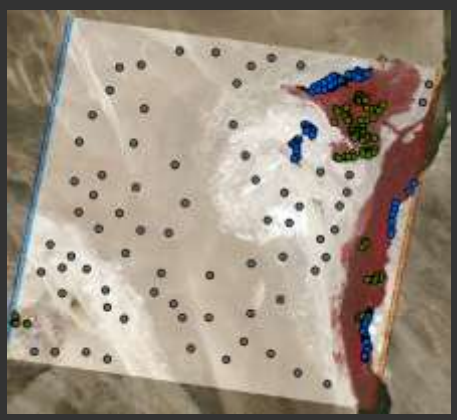
48

One technique that is generally useful for cleaning up classification results is sieving. Sieving methods identify small, contiguous areas of a class that are smaller than a certain area or number of pixels. These small areas are then removed and recoded to the surrounding class. For example, a small forest clearing that is smaller than the size threshold could be identified, removed, and recoded to the surrounding "forest" class.

Sieving operations are commonly used to minimize salt-and-pepper issues common to pixel-based classification. These methods can be used to generalize the results, make them more cartographically pleasing, and allow for the product to more closely approximate or honor the defined minimal mapping unit (MMU).

Example of Class Separability

Example 1

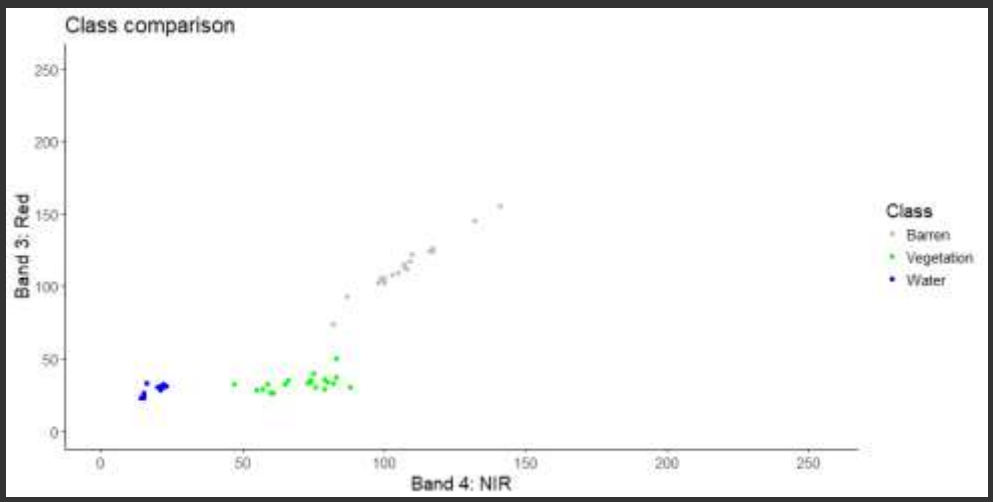


Landsat 5 TM October 18, 2003

50

To end this module, I have provided some examples to explore class separability. This first example is from a Landsat 5 Thematic Mapper (TM) scene covering a portion of the Nile River valley and adjacent desert collected on October 18, 2003. The points represent training samples for three classes: barren, vegetation, and water.

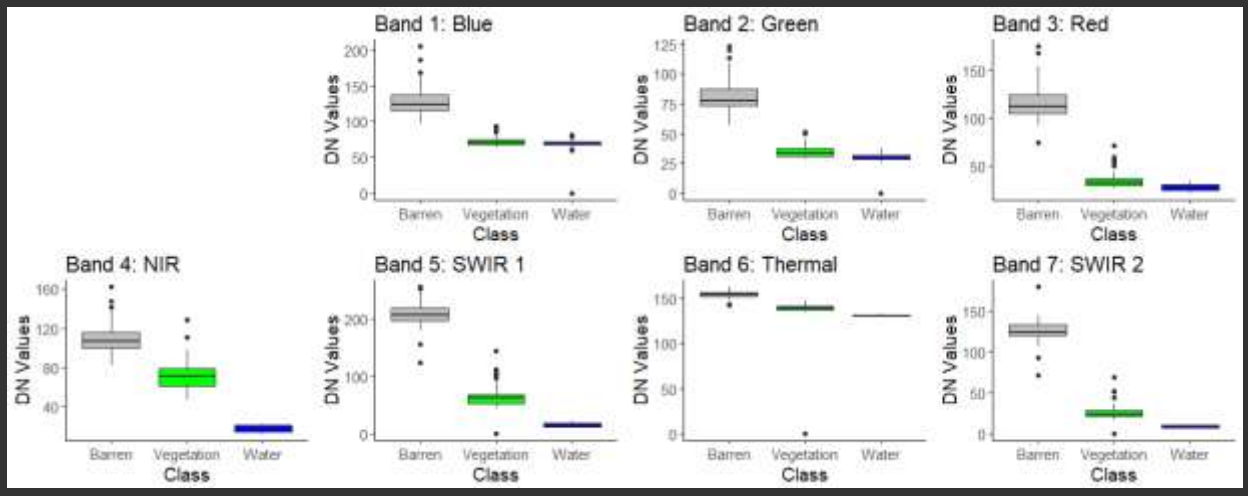
Class Comparison in Red and NIR Bands



51

This plot shows the distribution of the sample points in a space defined by the NIR and Red bands. Generally, these classes seem to be well separated using these two bands, as there is little overlap between the classes.

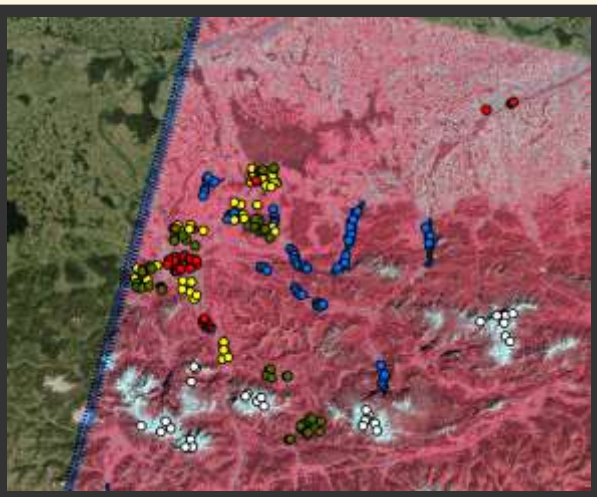
Class Comparison in All Bands



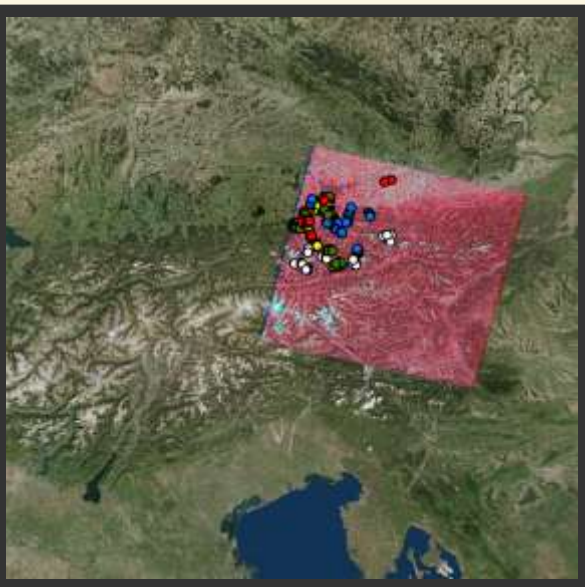
52

The provided boxplots show the distribution of the training samples for each TM spectral band by class. We see strong separation in the Red, NIR, and SWIR bands. The thermal band does not appear to differentiate the classes well.

Example 2



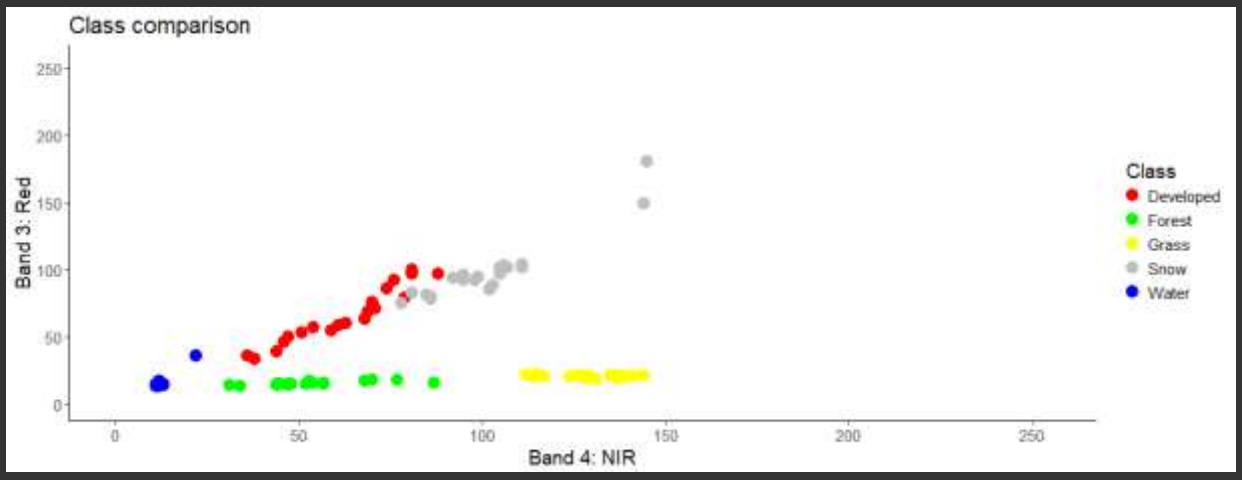
Landsat 5 TM August 23, 2011



53

This second example is another Landsat 5 TM image. It represents a section of the Bavarian and Austrian Alps. Training data for five classes have been collected: forest, grass, developed, snow, and water.

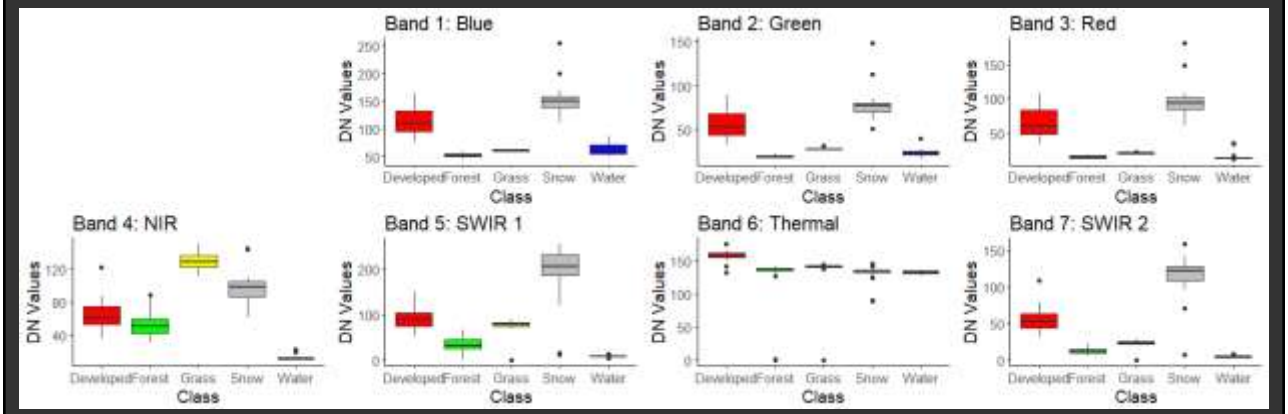
Class Comparison in Red and NIR Bands



54

This graph again shows the training samples relative to the NIR and Red bands. It appears that the forest, grass, and water classes are well differentiated. However, there is some overlap between the developed and snow classes.

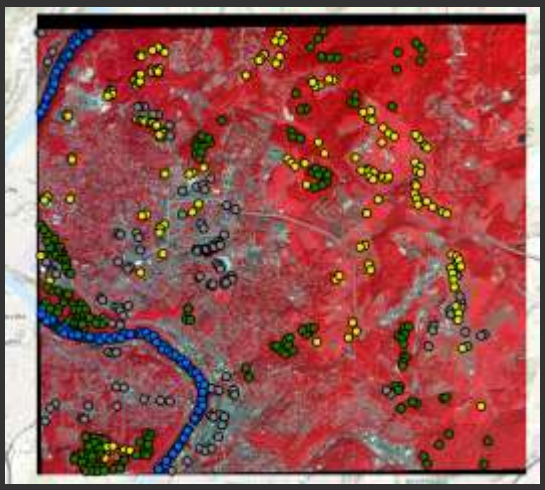
Class Comparison in All Bands



55

Assessing all the available bands, some bands seem to be more useful than others. The visible bands seem to differentiate the developed and snow classes from the other three classes. The NIR and SWIR data appear to be more useful for differentiate the forest and grassland classes. The thermal data do not seem to be of much value.

Example 3



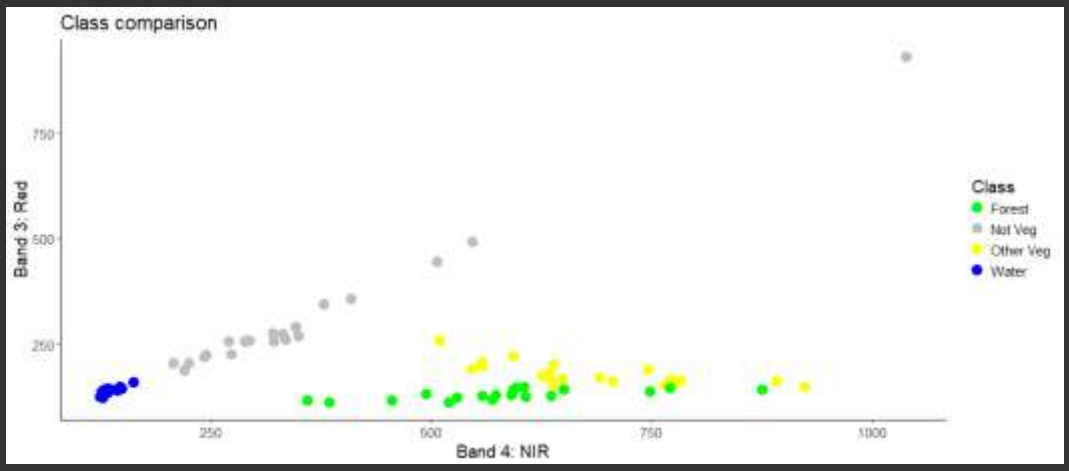
QuickBird Image of Morgantown

56

This last example uses a QuickBird image of Morgantown, WV. Compared to the Landsat data explored above, this sensor provides a higher spatial resolution, but the spectral resolution is reduced. Only four bands are provided: Blue, Green, Red, and NIR.

Training data have been created for four classes: barren, grass, forest, and water.

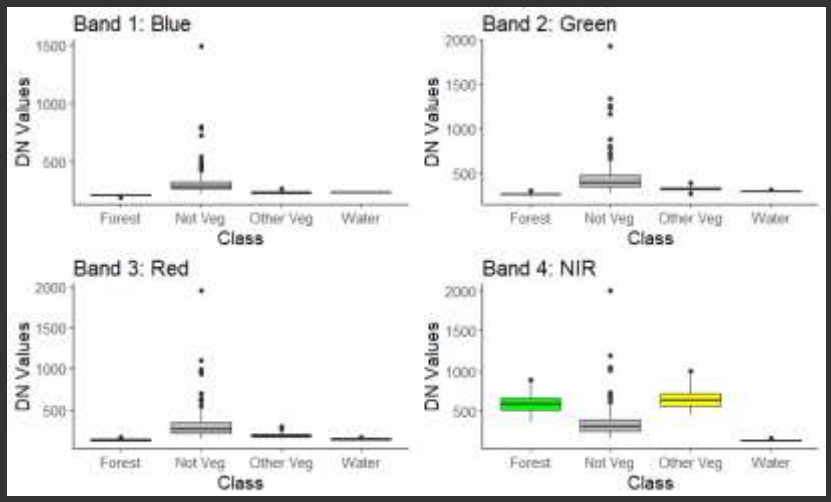
Class Comparison in Red and NIR Bands



57

Using the Red and NIR bands, it appears that water and barren areas can be well differentiated. However, overlap, or confusion, between forest and grass seems to be an issue.

Class Comparison in All Bands



58

Comparing the distribution of the classes across all four bands suggest that they are not well differentiated in the visible spectrum. Some enhanced differentiation is offered with the included NIR bands. However, some overlap is still observed, especially between forest and grass.

The lack of spectral resolution here, such as SWIR bands, may be an issue in this classification.

Result

59

Finally, this slide shows a classification result obtained using the QuickBird image and example training samples. Can we quantify the level of accuracy achieved and get a sense of what classes are most confused? This will be the topic of the next module, in which we investigate accuracy assessment techniques.

This is the end of this lecture module.

Please return to the West Virginia View Webpage for additional content.

60

Thanks! Hope you found this useful.