



Remotely Sensing Cities and Environments

Lecture 7: Classification The Big Questions (Lecture 6 continued) and Accuracy

28/06/2022 (updated: 07/03/2025)

✉ a.maclachlan@ucl.ac.uk

🌐 andymac.uk

🐦 andymaclachlan

👤 andrewmaclachlan

📍 Centre for Advanced Spatial Analysis, UCL

PDF PDF presentation

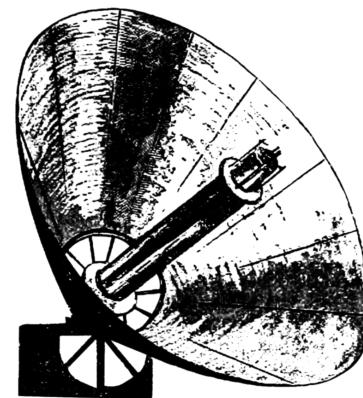
How to use the lectures

- Slides are made with `xaringan`
-  In the bottom left there is a search tool which will search all content of presentation
- Control + F will also search
- Press enter to move to the next result
-  In the top right let's you draw on the slides, although these aren't saved.
- Pressing the letter `o` (for overview) will allow you to see an overview of the whole presentation and go to a slide
- Alternatively just typing the slide number e.g. 10 on the website will take you to that slide
- Pressing alt+F will fit the slide to the screen, this is useful if you have resized the window and have another open - side by side.

Lecture outline

Part 1: Landcover classification
(continued)

Part 2: Accuracy



Source:Original from the British Library. Digitally enhanced by rawpixel.

What do we need (current or historic)
landcover data for?

Can we just used pre-classified data

Pre-classified data

- GlobeLand30 - 30m for 2000, 2010 and 2020:
http://www.globallandcover.com/home_en.html?type=data
- European Space Agency's (ESA) Climate Change Initiative (CCI) annual global land cover (300 m) (1992-2015): <https://climate.esa.int/en/projects/land-cover/data/>
- Dynamic World - near real time 10m: <https://www.dynamicworld.app/explore/>
 - A major benefit of an AI-powered approach is the model looks at an incoming Sentinel-2 satellite image and, for every pixel in the image, estimates the degree of tree cover, how built up a particular area is, or snow coverage if there's been a recent snowstorm, for example
- MODIS: <https://modis.gsfc.nasa.gov/data/dataproducts/mod12.php>
- Google building data: <https://sites.research.google/open-buildings/>

Dynamic World

Data

Training

Pre-processing

Normalisation

Classification

CNN

CNN 2

Accuracy

Example

Notes

Radiant MLHub

- Semi-supervised approach
 - Divided World into regions (Western Hemisphere (160°W to 20°W), Eastern Hemisphere-1 (20°W to 100°E), and Eastern Hemisphere-2 (100°E to 160°W))
 - Divided them into 14 Biomes
 - Stratified samples based on NASA MCD12Q1 land cover for 2017 + others

Before we progress....thoughts on this?

What was the data (SR, TOA)

How was it trained

What are the issues

Do you think it's any good

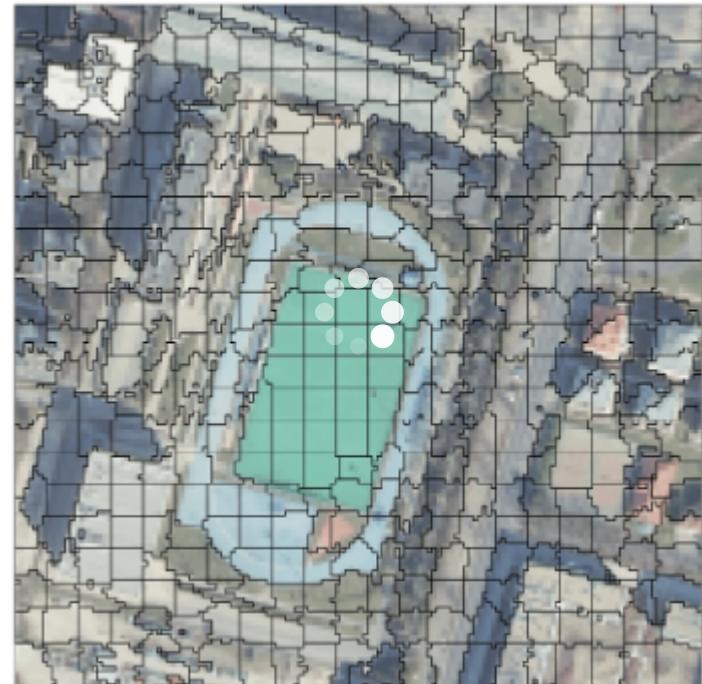
Next up

Object based image analysis and sub pixel
analysis

Object based image analysis (OBIA)

- Does a raster cell represent an object on the ground?
- Instead of considering cells we consider shapes based on the similarity (homogeneity) or difference (heterogeneity) of the cells = **superpixels**
- **SLIC** (Simple Linear Iterative Clustering) Algorithm for Superpixel generation is the **most common method**
 - regular points on the image
 - work out spatial distance (from point to centre of pixel) = **closeness to centre**
 - colour difference (RGB vs RGB to centre point) = **homogeneity of colours**

Iteration: 1



Supercells Source:Nowosad 2021

Object based image analysis (OBIA) 2

- Each iteration the centre moves- 4-10 is best (based on original paper)
- The values can change and the boarders move (like k-means?)
- Doesn't consider connectivity = very small cells
- Can enforce connectivity (remove small areas and merge them)
- S = distance between initial points
- m = compactness = balance between physical distance (larger value) and colour (spectral distance, then smaller m)

- Can only use Euclidean distance in SLIC



Supercells Source:[Darshite Jain](#)

Object based image analysis (OBIA) 3

- **Supercells** package can use any distance measure (e.g. dissimilarity)
- k = number of super pixels
- compactness = impact of spatial (higher value = squares) vs colour (lower value)
- transform = not on raw data, but to LAB colour space (Lab^* , is a color space defined by the International Commission on Illumination (abbreviated CIE))
- We can then take the **average values per object** and classify them using methods we've seen
- Other metrics can also be computed - e.g. length to width ratio (see Jensen p.418)



Supercells Source:Nowosad 2021

Object based image analysis (OBIA) 3

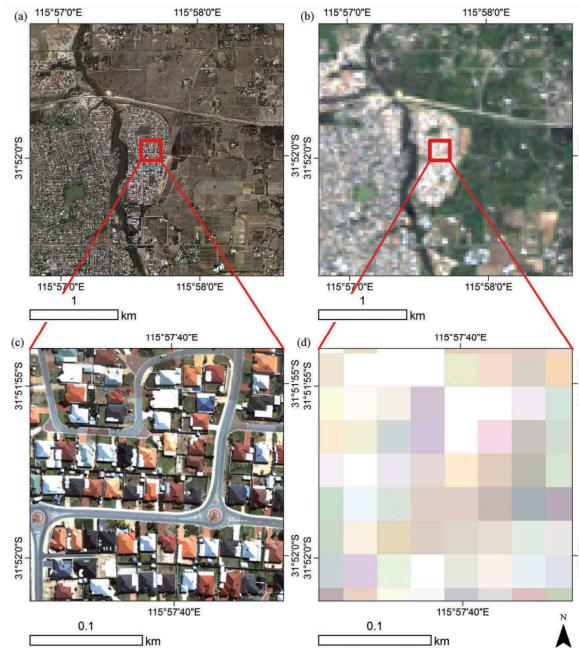
Note that there are many OBIA classifiers, they all do similar, but slightly different processes - see Jensen page 415

A more advanced package would be **SegOptim** that can use algorithms from other software



Sub pixel analysis

If you have a pixel composed of a range of land cover types should it be classified as one landcover or should we calculate the proportions ?

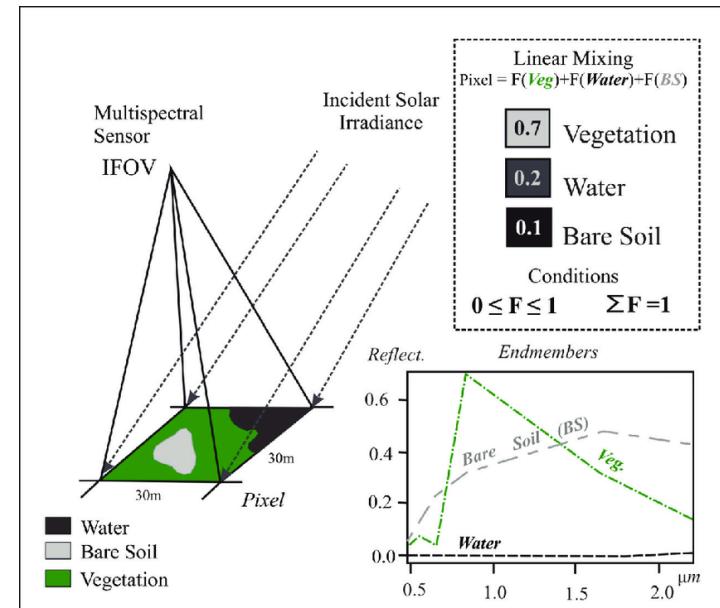


Comparison of true colour high spatial resolution data (a) (acquired from 14 March 2007) and Landsat surface reflectance (b) (acquired on 6 October 2007 [path 112]), highlighting the spatial detail captured by high-resolution imagery (c) and the same areas as observed by Landsat (d) for the subset East Beechboro used within this study Source:[MacLachlan et al. 2017](#)

Sub pixel analysis

Termed (all the same): Sub pixel classification, Spectral Mixture Analysis (SMA), Linear spectral unmixing

- SMA determines the **proportion** or **abundance** of landcover per pixel
- the assumption that reflectance measured at each pixel is represented by the linear sum of endmembers weighted by the associated endmember fraction
- Typically we have a few endmembers that are **spectrally pure**



Source: Machado and Small (2013) 2017

In R we can use MESMA from the package RStoolbox

Sub pixel analysis 2

- Sum of end member reflectance * fraction contribution to best-fit mixed spectrum

$$p_\lambda = \sum_{i=1}^n (p_{i\lambda} * f_i) + e_\lambda$$

p_λ = The pixel reflectance

$p_{i\lambda}$ = reflectance of endmember i

f_i = fractional cover of end member i

n = number of endmembers

e_λ = model error

See, Jensen page 480 - following example taken from there

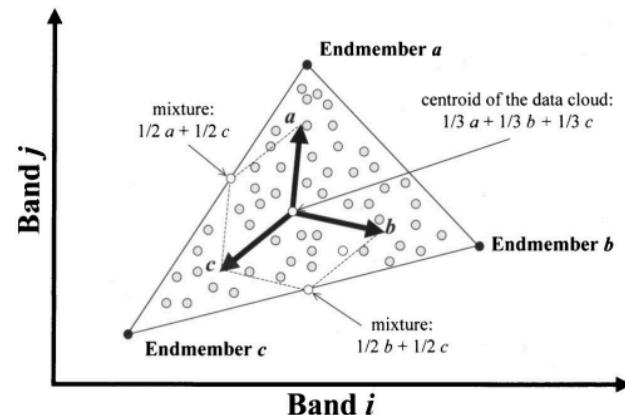


Fig. 1. Scatterplot of two-dimensional spectral data illustrating the physical interpretation of a mixture model based on endmembers.

Source: Plaza et al. (2002)

Sub pixel analysis 3

Not as complicated as it looks...here are some end members for bands 3 and 4

Band	Water	Vegetation	Soil
3	13	22	70
4	5	80	60

left here is the pixel we are trying to model = end members * the fraction (proportion)

$$\begin{bmatrix} \text{band3} \\ \text{band4} \\ \text{sumto1} \end{bmatrix} = \begin{bmatrix} \text{pwater} & \text{pveg} & \text{psoil} \\ \text{pwater} & \text{pveg} & \text{psoil} \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \text{fwater} \\ \text{fveg} \\ \text{fsoil} \end{bmatrix}$$

We take the **inverse matrix (\wedge^{-1})** of endmembers ...

$$\begin{bmatrix} 13 & 22 & 70 \\ 5 & 80 & 60 \\ 1 & 1 & 1 \end{bmatrix} \text{to...} \begin{bmatrix} -0.0053 & -0.0127 & 1.1322 \\ -0.0145 & 0.0150 & 0.1137 \\ 0.0198 & -0.0024 & -0.2460 \end{bmatrix}$$

Sub pixel analysis 4

Then solve ... if our values for the pixel are **25** (band 3) and **57** (band 4) the rows of the first matrix are multiplied by the columns of the second one

$$\begin{bmatrix} f_{\text{water}} \\ f_{\text{veg}} \\ f_{\text{soil}} \end{bmatrix} = \begin{bmatrix} -0.0053 & -0.0127 & 1.1322 \\ -0.0145 & 0.0150 & 0.1137 \\ 0.0198 & -0.0024 & -0.2460 \end{bmatrix} \begin{bmatrix} 25 \\ 57 \\ 1 \end{bmatrix}$$

This looks like...(from **matrix calculator**)

$$\begin{pmatrix} \frac{-53}{10000} * 25 + \frac{-127}{10000} * 57 + \frac{5661}{5000} * 1 \\ \frac{-29}{2000} * 25 + \frac{3}{200} * 57 + \frac{1137}{10000} * 1 \\ \frac{99}{5000} * 25 + \frac{-3}{1250} * 57 + \frac{-123}{500} * 1 \end{pmatrix}$$

- Different denominators are just for the different values

Sub pixel analysis 5

And gives...

$$\begin{bmatrix} 0.27 \\ 0.61 \\ 0.11 \end{bmatrix} = \begin{bmatrix} -0.0053 & -0.0127 & 1.1322 \\ -0.0145 & 0.0150 & 0.1137 \\ 0.0198 & -0.0024 & -0.2460 \end{bmatrix} \begin{bmatrix} 25 \\ 57 \\ 1 \end{bmatrix}$$

This means that within this pixel we have:

- 27% water
- 61% vegetation
- 11% soil

Sub pixel analysis 6

Issues / considerations:

- Pixel purity ?
- Number of End members
 - simplify the process and use the **V-I-S model** in urban areas:
Vegetation-Impervious surface-Soil (V-I-S) fractions
- Multiple endmember spectral analysis (MESMA)
 - Increase computation
 - or use a spectral library

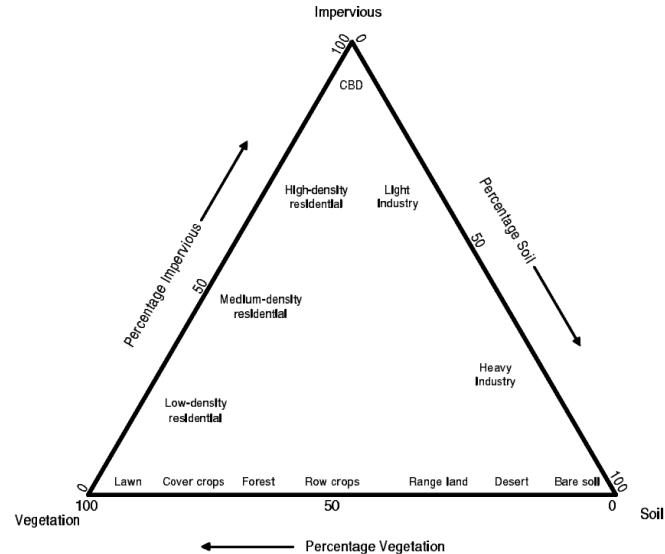


Figure 1. VIS—Vegetation–Impervious surface–Soil—model of Ridd (1995).

Source: Phinn et al. (2002) Monitoring the composition of urban environments based on the vegetation-impervious surface-soil (VIS) model by subpixel analysis techniques,

Accuracy assessment

After producing and output we need to assign an accuracy value to it (common to machine learning).

In remote sensing we focus on:

- PA Producer accuracy (recall or true positive rate or sensitivity) - vertical **the classification results meet the expectation of the creator**
- UA User's accuracy (consumer's accuracy or precision or positive predictive value) - horizontal **pixels are incorrectly classified as a known class when they should have been classified as something else**
- OA the (overall) accuracy

		Actual			
		Class C	Not Class C		
Classified	Class C	TP 510	FP 228	UA 69.1%	CE 30.9%
	Not Class C	FN 111	TN 4624	NPV 97.7%	FOR 2.3%
		PA 82.1%	TNR 95.3%		
		OE 17.9%	FPR 4.7%	OA 93.8%	

Table 1. Binary confusion matrix (Example data are taken from Campbell, 1996)

Source: Barsi et al. 2018 Accuracy Dimensions in Remote Sensing

Accuracy assessment 2

		Actual			
		Class C	Not Class C		
Classified	Class C	TP 510	FP 228	UA 69.1%	CE 30.9%
	Not Class C	FN 111	TN 4624	NPV 97.7%	FOR 2.3%
		PA 82.1%	TNR 95.3%	OA 93.8%	
		OE 17.9%	FPR 4.7%		

Table 1. Binary confusion matrix (Example data are taken from Campbell, 1996)

Source: Barsi et al. 2018 Accuracy Dimensions in Remote Sensing

Where model is correct

- True positive = model predicts positive class correctly
- True negative = model predicts negative class correctly

Where model is incorrect

- False positive = model predicts positive, but it is negative
- False negative = model predicts negative, but it is positive

Accuracy assessment 3

- **producer's accuracy** defined as the fraction of correctly classified pixels (TP) compared to ground truth data (TP+FN) $\frac{TP}{TP+FN}$
- **user's accuracy** defined as the fraction of correctly classified pixels (TP) relative to all others classified as a particular land cover(TP+FP) $\frac{TP}{TP+FP}$ - FP is different
- **overall accuracy** that represents the combined fraction of correctly classified pixels (TP +TN) across all land cover types (TP+FP+FN+TN) $\frac{TP+TN}{TP+FP+FN+TN}$

		Actual			
		Class C	Not Class C		
Classified	Class C	TP 510	FP 228	UA 69.1%	CE 30.9%
	Not Class C	FN 111	TN 4624	NPV 97.7%	FOR 2.3%
		PA 82.1%	TNR 95.3%	OA 93.8%	
		OE 17.9%	FPR 4.7%		

Table 1. Binary confusion matrix (Example data are taken from Campbell, 1996)

Source: Barsi et al. 2018 Accuracy Dimensions in Remote Sensing

Accuracy assessment 4

Example

	Expert Majority										
		Water	Trees	Grass	Flooded Vegetation	Crops	Shrub & Scrub	Built Area	Bare Ground	Snow/Ice	Precision/ User's:
Dynamic World	Water	7801513	49529	39511	187484	511166	59061	46461	197843	4178	87.70%
	Trees	127510	20225220	1280463	963384	2835774	3215029	293635	150630	8956	69.50%
	Grass	6465	135888	1415436	170294	1917310	405560	41812	151648	695	33.30%
	Flooded Veg.	54365	73337	56852	764482	133162	72474	6878	40949	6	63.60%
	Crops	23790	144438	261750	36363	10821384	707154	134486	328916	38	86.90%
	Shrub & Scrub	18649	1034643	551074	99947	1476117	4498630	157166	708463	32079	52.50%
	Built Area	11187	156117	57647	3935	666712	104009	6620303	82974	744	85.90%
	Bare Ground	178304	16203	40275	8835	400529	1022049	198274	2722023	48658	58.70%
	Snow & Ice	68319	199018	8656	550	59786	109192	14527	214993	1422556	67.80%
	Recall/Producer's:	94.10%	91.80%	38.10%	34.20%	57.50%	44.10%	88.10%	59.20%	93.70%	71.30%

Table 8. Confusion matrix of Dynamic World to Expert Majority, i.e. valid where, amongst labels, there was consensus or only one expert labeled (n = 78,916,422).

Source: Brown et al. 2022 Dynamic World, Near real-time global 10m land use land cover mapping

Accuracy assessment 5

- Errors of omission (100-producer's accuracy)
 - Landcover omitted from correct class
 - Type 1 error
 - $\text{Urban} = (22/22+1) \frac{\text{TP}}{\text{TP}+\text{FN}}$
 - Urban producer = $22/23$, 95.65%
- Errors of commission (100- user's accuracy)
 - Classified sites for incorrect classifications
 - $\text{Urban user} = (22/22+9)$, $70.97\% \frac{\text{TP}}{\text{TP}+\text{FP}}$
- Kappa coefficient

		Reference Data			
		Water	Forest	Urban	Total
Classified Data	Water	21	6	0	27
	Forest	5	31	1	37
	Urban	7	2	22	31
	Total	33	39	23	95

Source: [Earth Systems Science and Remote Sensing](#)

Accuracy assessment 6

Producer accuracy ...

i am pleased that 95.65 % of the urban area that was identified in the reference is urban in the classification

User accuracy ...

as a user i only find that only 70.97% of the time when i visit an urban area is it acutally urban

Overall accuracy is 77.89%

- is this acceptable for the user?
- there is no single right choice for accuracy measurements

		Reference Data			
		Water	Forest	Urban	Total
Classified Data	Water	21	6	0	27
	Forest	5	31	1	37
	Urban	7	2	22	31
	Total	33	39	23	95

Source:[Earth Systems Science and Remote Sensing](#)

- This can also be changed to a fuzzy matrix (e.g. Deciduous forest classified as evergreen forest - see Jensen p.575)

Accuracy assessment 7

To Kappa or not to Kappa?

- Designed to express the accuracy of an image compared to the results by chance
- Ranges from 0 to 1

"Sadly the calls to abandon the use of the kappa coefficient in accuracy assessment seem to have fallen on deaf ears. It may be that the kappa coefficient is still widely used because it has become ingrained in practice and there may be a sense of obligation to use it"

$$k = \frac{p_o - p_e}{1 - p_e}$$

p_o is the proportion of cases correctly classified (accuracy) $\frac{TP+TN}{TP+FP+FN+TN}$

p_e expected cases correctly classified by chance (further equations in [Foody 2020](#)) or [Earth Systems Science and Remote Sensing](#)

Source: Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. Foody 2020

Kappa example

Accuracy Assessment: <i>Kappa</i>						
$\hat{K} = 0.46$						
Class types determined from classified map	Class types determined from reference source					User's Accuracy
	# Plots	Conifer	Hardwood	Water	Totals	
	Conifer	50	5	2	57	88%
	Hardwood	14	13	0	27	48%
	Water	3	5	8	16	50%
	Totals	67	23	10	100	
Producer's Accuracy		75%	57%	80%		Total: 71%

$$Po = 0.71,$$

$$Pe = (67*57+27*23+10*16)/10000 = 0.46$$

$$K = (0.71 - 0.46) / (1 - 0.46) = 0.46296296296$$

Earth Systems Science and Remote Sensing(<https://medium.com/@wenzhao.li1989/accuracy-assessment-d164e492274b>)

Kappa issues

- What is a good value?

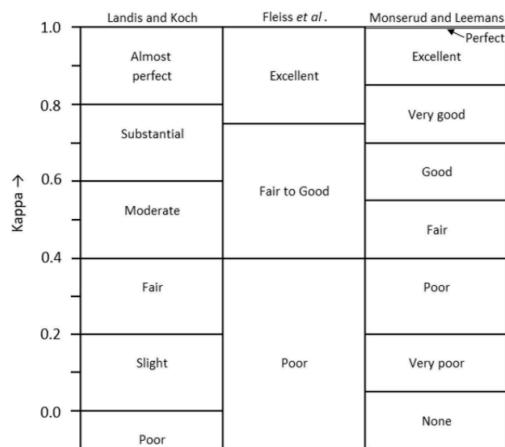


Fig. 3. Three scales for the interpretation of the kappa coefficient (adapted and updated from Caplewski, 1994). The scales are those provided by Landis and Koch (1977, page 165); Fleiss et al. (2013, page 604) and Monserud and Leemans (1992, page 285). Note that the full scale of measurement does extend to -1 but the focus is usually on positive values only.

6

- How Kappa values can we have for different levels of accuracy (on x axis)

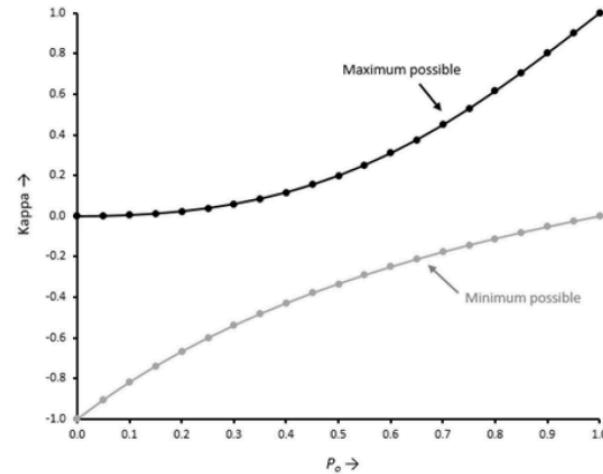


Fig. 4. Relationships between the maximum and minimum possible kappa coefficient with overall accuracy (p_o).

Source: Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. Foody 2020

Have i used Kappa?

See Jensen page 570

In remote sensing this is typically where
we'd stop...but not necessarily in machine
learning

A brief overview...

Beyond remote sensing

Beyond **traditional** remote sensing accuracy assessment...

Problem with recall (Producer accuracy)
vs Precision (User accuracy)

False positives (Producer) or false
negatives (User) more important?

- model with high recall (Producer accuracy) = true positives but some false positives (predicted urban but land cover that isn't urban)
- Model with high precision (User's accuracy) = actual urban but predicted other landcover
- See **MLU-explain** for an interactive example and next slide...

		Actual			
		Class C	Not Class C		
Classified	Class C	TP 510	FP 228	UA 69.1%	CE 30.9%
	Not Class C	FN 111	TN 4624	NPV 97.7%	FOR 2.3%
		PA 82.1%	TNR 95.3%	OA 93.8%	
		OE 17.9%	FPR 4.7%		

Table 1. Binary confusion matrix (Example data are taken from Campbell, 1996)

Source: Barsi et al. 2018 Accuracy Dimensions in Remote Sensing

In the next few slides we will change a decision threshold to see the effects on user's and producer's accuracy.

Focus on the changing decision threshold...

Our data is not balanced....

We can't have both a high producer accuracy (recall) and a high user's accuracy (precision)

- **user's accuracy (precision)** ratio of correctly predicted positive classes (TP) to all items predicted to be positive (TP+FP) $\frac{TP}{TP+FP}$

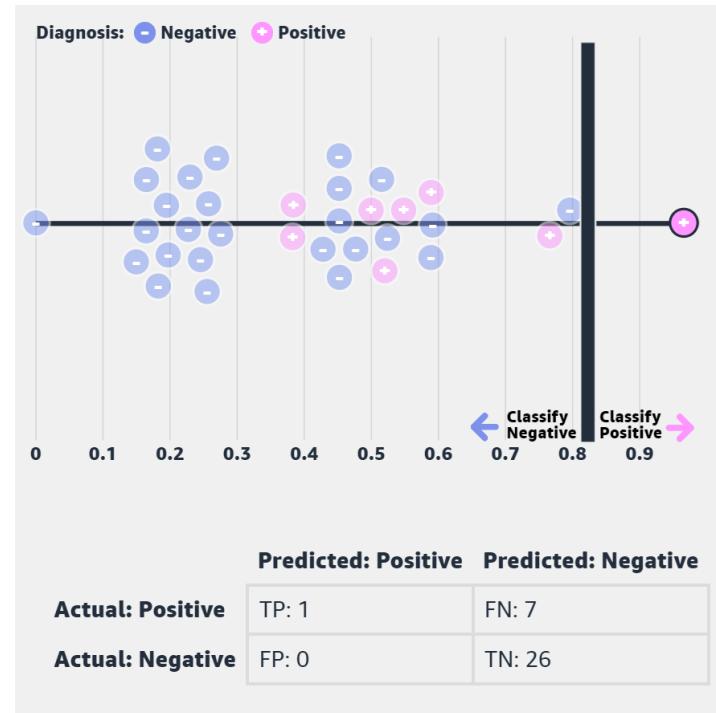
1/1=100%

or

8/19=42% (this is on the next slide)

how precise the model is at positive predictions

as a user i only find that only x% of the time when i visit a positive point it is actually positive...



Source:MLU-explain

Our data is not balanced...2

We can't have both a high producer accuracy (recall) and a high user's accuracy (precision)

- producer's accuracy (recall) the ratio of correctly predicted positive classes (TP) to all items that are actually positive (TP+FN) $\frac{TP}{TP+FN}$

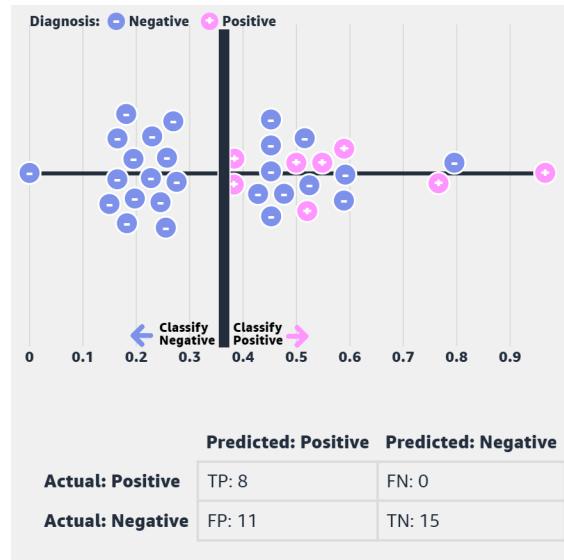
1/8=13% (from previous slide)

or

8/8=100%

how many positive points are correct

i am pleased that [as a producer] x% of points are correct compared to the reference



Source:MLU-explain

Our data is not balanced....3

user's accuracy (precision)

- I have gone to a site, the model predicted it to be urban, it is not urban...
- How well can the user use the data / classification

producer's accuracy (recall)

- I have gone all the urban sites, they were urban. **BUT** i can see in the distance a site that was predicted to be GRASS but is actually URBAN
- How well did the producer make the data/ classification

Combine them both...into a F1 score

F1



The F1-Score (or F Measure) combines both recall (Producer accuracy) and Precision (User accuracy):

- $$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Which equals....

- $$F1 = \frac{\text{TP}}{\text{TP} + \frac{1}{2} * (\text{FP} + \text{FN})}$$
- Value from 0 to 1, where 1 is better performance

Source:[MLU-EXPLAIN](#)

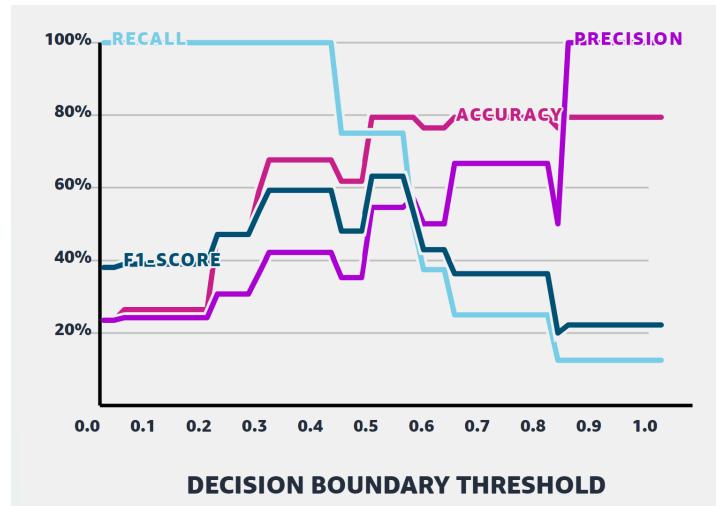
		Actual			
		Class C	Not Class C		
Classified	Class C	TP 510	FP 228	UA 69.1%	CE 30.9%
	Not Class C	FN 111	TN 4624	NPV 97.7%	FOR 2.3%
		PA 82.1%	TNR 95.3%	OA 93.8%	
		OE 17.9%	FPR 4.7%		

Table 1. Binary confusion matrix (Example data are taken from Campbell, 1996)

Source:[Barsi et al. 2018 Accuracy Dimensions in Remote Sensing](#)

F1 issues

- No True Negatives (TN) in the equation
 - negative categories that are correctly classified as negative
- Are precision and recall equally important?
 - precision (producer): how many positive points are correct
 - recall (user): how precise the model is at positive predictions
- What if our data is very unbalanced?
 - More negatives than positives?

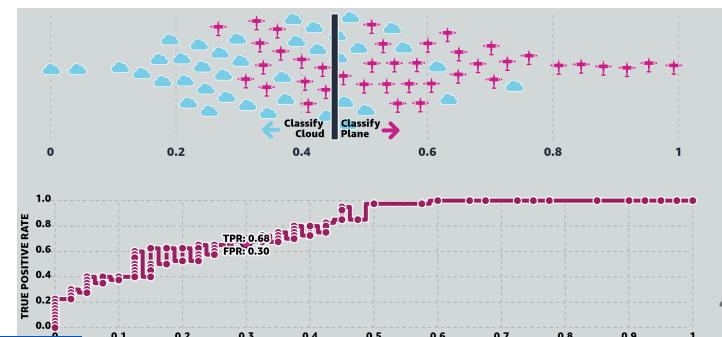


Source: [MLU-EXPLAIN](#)

Receiver Operating Characteristic Curve

Receiver Operating Characteristic Curve (the ROC Curve)

- Receiver Operating Characteristic Curve (the ROC Curve)
- Originates from WW2, USA wanted to minimize noise from radar to identify (true positives) and not miss aircraft...minimizing false positives (clouds)
- **Changing the threshold value of classifier** will change the True Positive rate
 - probability that a positive sample is correctly predicted in the positive class...planes predicted to be planes
- False positive rate: The probability that a negative sample is incorrectly predicted in the positive class...predicted planes...but are clouds
- Maximise true positives (1) and minimise false positives (0)
- plot starts from right - threshold 0 then TPR and FPR both 1 (right of the plot)



Receiver Operating Characteristic Curve

vertical columns here - uses whole matrix

- First is True positive
 - True positive rate = $TP/TP+FN$
- Second is False positive rate
 - False positive rate = $FP/FP+TN$

		Actual			
		Class C	Not Class C		
Classified	Class C	TP 510	FP 228	UA 69.1%	CE 30.9%
	Not Class C	FN 111	TN 4624	NPV 97.7%	FOR 2.3%
		PA 82.1%	TNR 95.3%	OA 93.8%	
		OE 17.9%	FPR 4.7%		

Table 1. Binary confusion matrix (Example data are taken from Campbell, 1996)

Source: Barsi et al. 2018 Accuracy Dimensions in Remote Sensing

Receiver Operating Characteristic Curve

Receiver Operating Characteristic Curve (the ROC Curve)

- True positive rate = good = every plane is a plane ?
- False positive rate = good = every cloud is predicted as noise (not a plane) ?

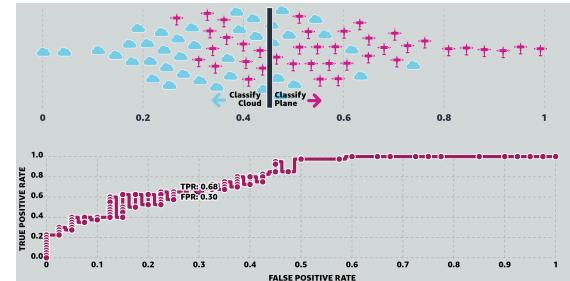
when threshold is 0 (everything classified as plane)

all planes are planes = TPR =
1

but ...

all clouds are planes = FPR =
1

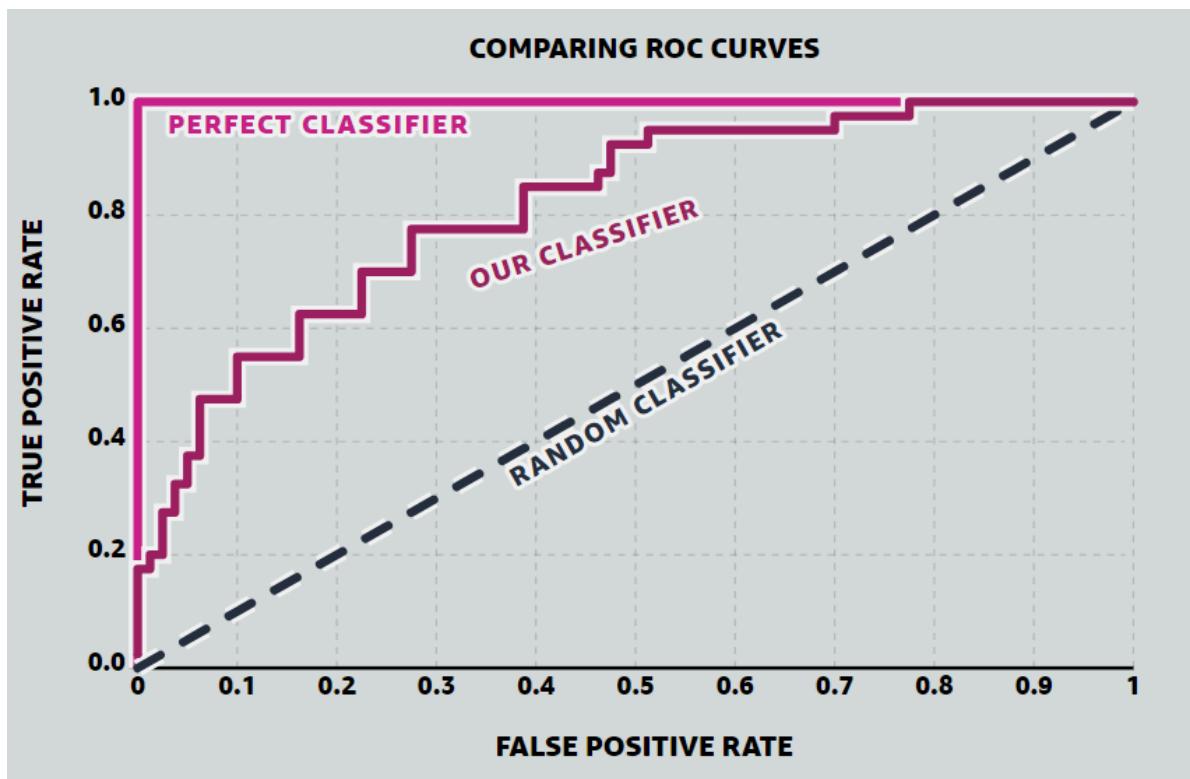
- Maximise true positives (1) and minimise false positives (0)



Source: [MLU-EXPLAIN](#)

Receiver Operating Characteristic Curve

Goal: Maximise true positives and minimise false positives... (e.g. urban vs other landcover)



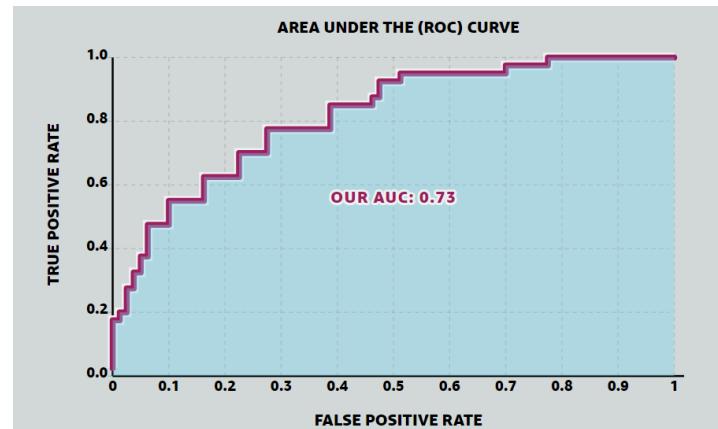
Area Under the ROC Curve

Area Under the ROC Curve (AUC, or AUROC)

- Simply the area under the curve
- Compare models easily (no need to look at the ROC curve)
- Perfect value will be 1, random will be 0.5

"The AUC is the probability that the model will rank a randomly chosen positive example more highly than a randomly chosen negative example..."

e.g. model always give positive from true negative (so always wrong) = AUC 0



Source:MLU-EXPLAIN

- used to determine how well your **binary** classifier stands up to a **gold standard binary classifier**. There is also a multi-class method that follows one vs one or one vs rest discussed last time

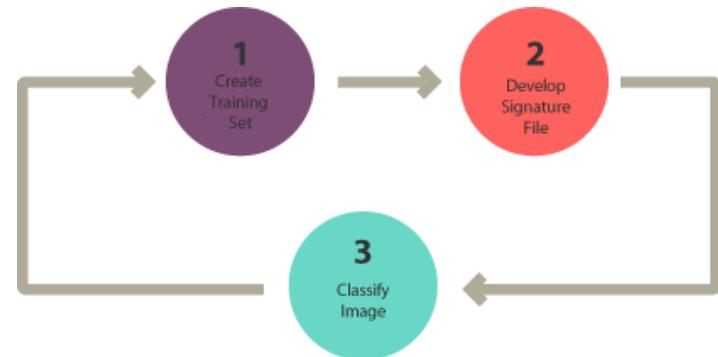
Next time (?)

How do we get test data for the accuracy assessment?

Remote sensing approach (sometimes)

Same process for all:

- class definition
- pre-processing
- training
- pixel assignment
- accuracy assessment



Source: [GIS Geography](#)

- Collect training data - suggested as around 250 pixels per class (Foody and Mather, 2006)
- Simply go and collect (or use Google Earth) ground truth data - 50 per class (Congalton, 2001).
- Produce an error matrix

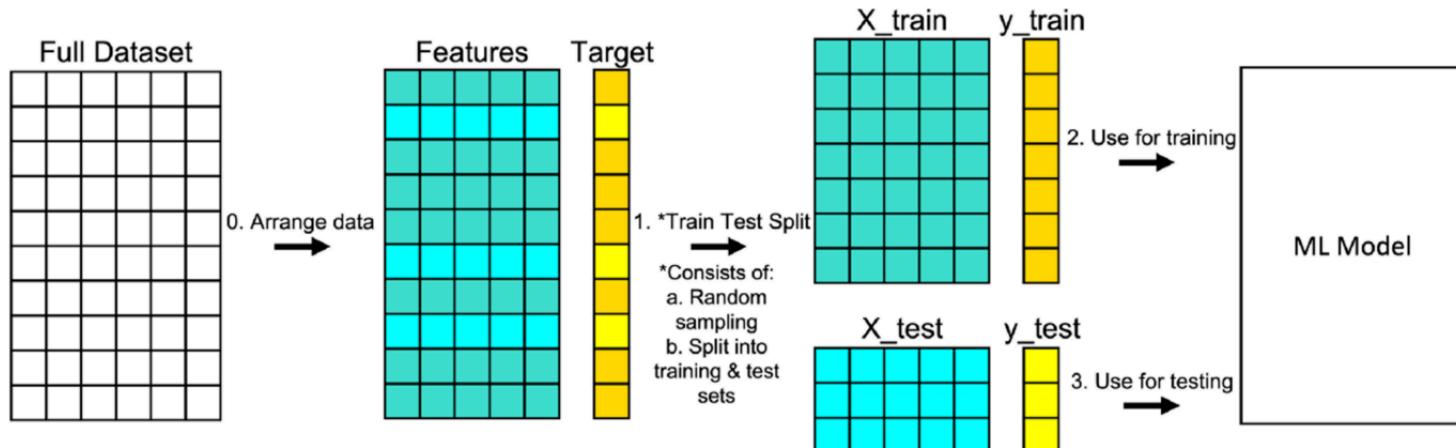
- You would need to consider a **sampling strategy**
 - Random sampling
 - Systematic sampling
 - Stratified sampling
 - Jensen p. 565

Problems?

Chapter 13 in Jensen (p.557) cover accuracy assessment...but not the following

Good approach - train and test split

- This is simply holding back a % of the original data used to train the model to then test it at the end
- See the [validation section \(10.6.7\)](#) for an example in linear regression



Source: Michael Galarnyk

Best approach - cross validation

Really classification of imagery is a machine learning task...

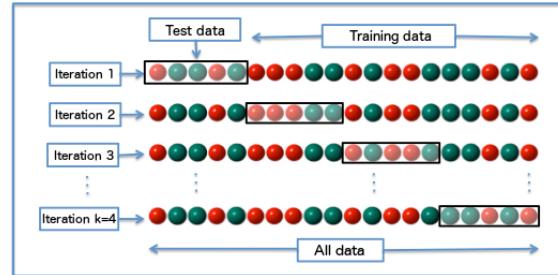
..So why can't we apply the same methods?

Perhaps as it is meant to be iterative...

...e.g. the classifier underpredicts urban then you can go and adjust the training data...

We might take the mean accuracy from the cross validation.

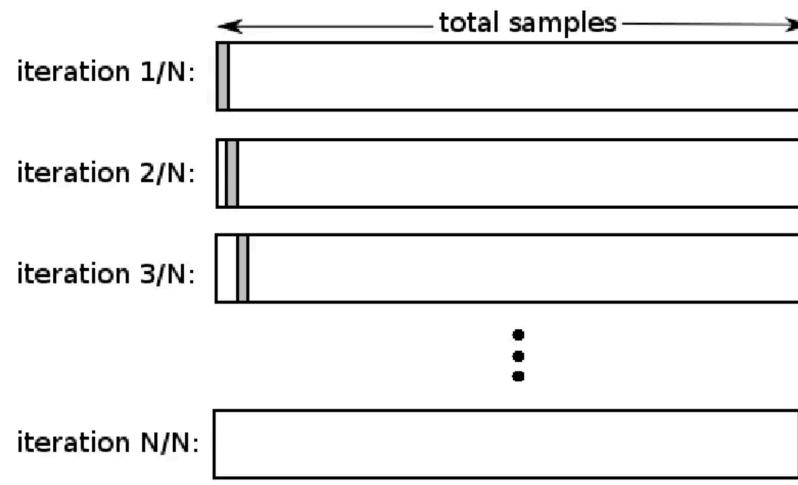
Leave one out cross validation is an extreme version where the folds (often 10) equals the number of samples in the data minus 1...next slide...



Source: [Wikipedia](#)

Leave one out cross validation

- An extreme version of cross validation
- Only for smaller datasets
- Uses all the data except 1, this 1 is the testing set
- Repeats though all of it



representation of leave one out cross validation

BUT..."Spatial autocorrelation between
training and test sets"

Remember spatial autocorrelation?

A measure of similarity between nearby data...

Best approach - cross validation

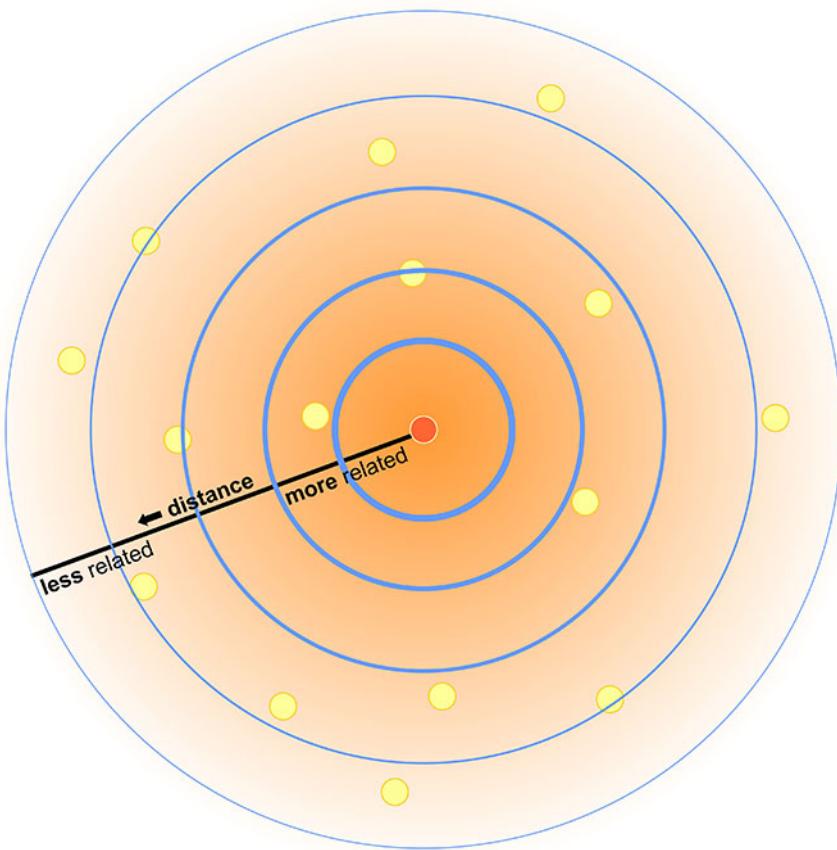
Waldo Tobler's first Law of Geography...

"everything is related to everything else, but near things are more related than distant things."

- Are training and testing points too close in geographic space?
- How can we deal with taking a sample of training data for testing when they are possibly from the same polygon of training data...

'Training' observations near the 'test' observations can provide a kind of 'sneak preview': information that should be unavailable to the training dataset.

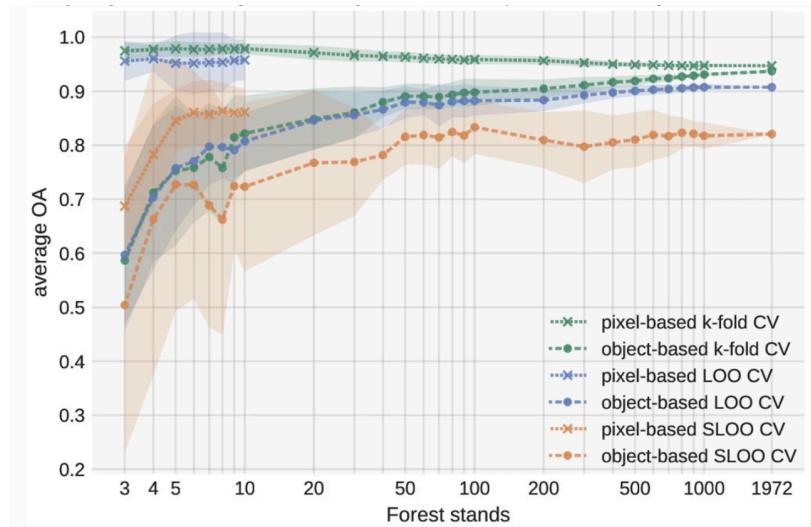
Source:[Lovelace et al. 2022](#)



Source: [Spatial is Special](#)

Spatial dependence....

Karasiak et al. 2022, Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing



Average overall accuracy based on the RF classifier for each cross-validation strategy (k-fold CV, LOO CV, SLOO CV) at pixel and object levels. Models were fitted with reference samples of Herault-34 and repeated 10 times (i.e. the y-axis provides the average OA value \pm standard deviation). The premature stopping of the pixel-based LOO and SLOO CV approaches was due to excessive computational time. Source: Karasiak et al. 2022

- (1) a k-fold cross-validation (k-fold-CV) based on random splitting
- (2) a non-spatial leave-one-out cross-validation (LOO CV)
- (3) a spatial leave-one-out cross-validation (SLOO CV) using a distance-based buffer relying on Moran's I statistics.

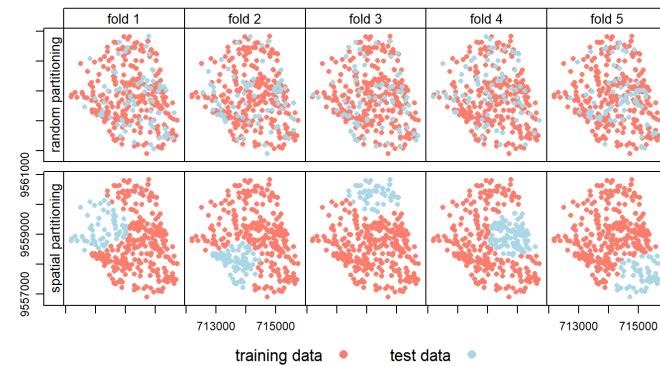
Why do the pixels perform better than the objects?

Spatial cross validation

Spatial cross validation

- spatially partition the folded data, folds are from cross validation
- disjoint (no common boundary) using k -means clustering (number of points and a distance)
- same as cross validation but with clustering to the folds...
- stops our training data and testing data being near each other...

in other words this makes sure all the points (or pixels) we train the model with are far away from the points (or pixels) we test the model with



Spatial visualization of selected test and training observations for cross-validation of one repetition. Random (upper row) and spatial partitioning (lower row).
Source: Lovelace et al. 2022

How does this compare to ideas we saw in CASA0005
(e.g. lag, error, GWR)

Did we test our models there

How did we deal with nearby points being related ?

Did we need to generalise for new data or not ?

If not then we don't need to test the model

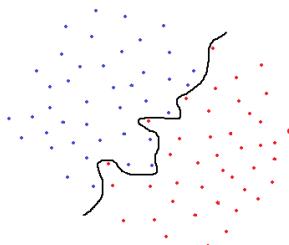
Spatial cross validation 2

Lovelace et al. (2022) use a Support Vector Machine classifier that requires hyperparameters (set before the classification)

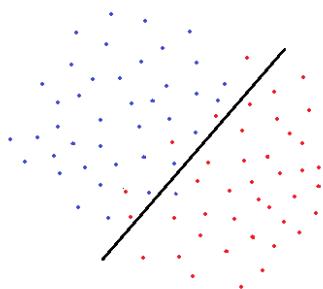
this is on the training data (e.g. band 4 between grass and urban land cover)

Standard SVM then the classifier will try to **overfit** = perfect for the current data but useless for anything else...

Cortes and Vapnik - **soft margin**, permit misclassifications = controlled with **C**



Source:Soner Yildirim



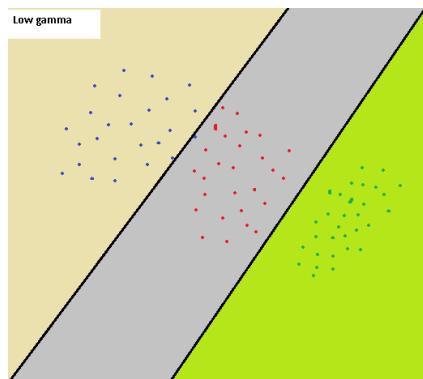
Source:Soner Yildirim

- **C** = adds penalty (proportional to distance from decision line) for each classified point. Small = image on right, large = image on left. **changes the slope**

Spatial cross validation 3

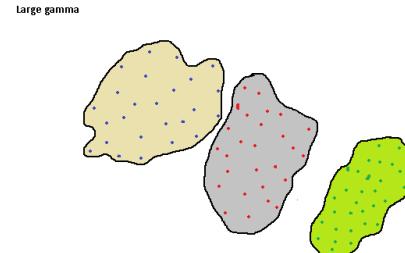
Lovelace et al. (2022) use a Support Vector Machine classifier that requires hyperparameters (set before the classification)

- **Gamma (or also called Sigma)** = controls the influence of a training point within the classified data
 - low = big radius and many points in same group
 - high = low radius and many groups



Source:Soner Yildirim

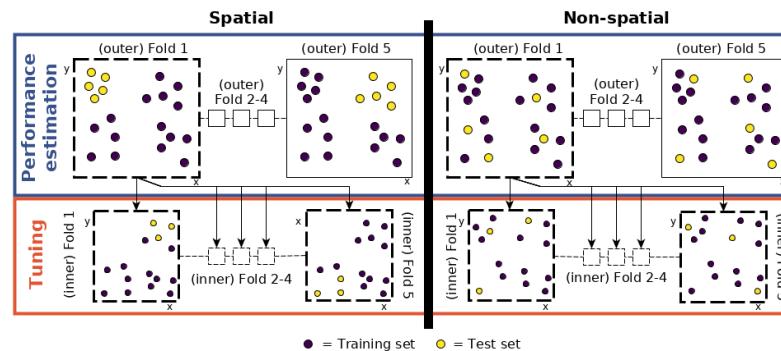
Source:Soner Yildirim



Source:Soner Yildirim

Spatial cross validation 4

- **Performance level** each spatial fold (taken from our first k-means cross validation fold division). = Top row below, a typical cross validation fold...
- **Tuning level** each fold (outer) is then divided into 5 again (inner fold).= Bottom row below. Test data is from the fold, not instance test data.
- **Performance estimation** Use the 50 randomly selected hyperparameters in each of these inner subfolds, i.e., fit 250 models with random C and Gamma use the best values to outer fold, based on **AUROC** with inner fold testing data. Note, this is not the main testing data.



Schematic of hyperparameter tuning and performance estimation levels in CV. (Figure was taken from Schratz et al. (2019).. Source:Lovelace et al. 2022

See the figure on the previous slide..."Using the same data for the performance assessment and the tuning would potentially lead to overoptimistic results"

...this means tuning on a "normal cross validation fold" is not representative...

Here tuning of parameters is made on a different subset of the data within each fold...

Spatial cross validation 6

Here we have...

- 1 outer fold has 5 inner folds with 50 randomly selected hyper parameters = 250 models for **C** and **Gamma**

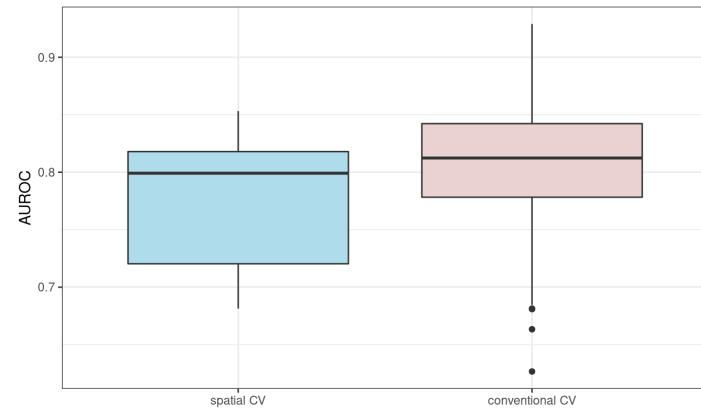
It's 5 over folds

- Each repetition = 1,250 ($250 * 5$)

It's 100 times repeated (5 fold cross validation)

- 125,500 models for best hyperparameters

So what



Boxplot showing the difference in GLM AUROC values on spatial and conventional 100-repeated 5-fold cross-validation. Source: [Lovelace et al. 2022](#)

Question: What happens if a classification model doesn't consider spatial autocorrelation?

The model will have better accuracy than it actually does

Question: What methods can we use to deal with it?

Summary

- How should we consider / process EO data, **objects, pixels, mixed pixel, mixed objects ?**
- What data should we use assess the accuracy of our classification models
 - New dataset to test the output with
 - Train / split the training data
 - Cross validation
- When we have a test dataset **how do we assess the accuracy**
 - Error matrix
 - Kappa
- When training and testing our classification models do we need to **consider spatial autocorrelation?** Do the following help
 - Object based image analysis
 - Spatial cross validation

Reading

Land Use Cover Datasets and Validation Tools

<https://doodles.mountainmath.ca/blog/2019/10/07/spatial-autocorrelation-co/>

<https://geocompr.robinlovelace.net/spatial-cv.html>

<https://link.springer.com/article/10.1007/s10994-021-05972-1>

<https://machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/#:~:text=fold%20Cross%2DValidation-,Leave%2Done%2Dout%20cross%2Dvalidat>

<https://machinelearningmastery.com/k-fold-cross-validation/>

