
Probabilistic Classifiers Using Nearest Neighbor Balls

Climate Change *Workshop, Malta, March, 2009*

Bo Ranney & Jun Yu

Centre of Biostochastics

Swedish University of Agricultural Sciences

- **Climate change implies e.g. new species composition in the forest**
- **For assessment of ecosystem health and quantification of spatial and temporal variation in land use and landscape patterns.**
- **Remote sensing should be a useful tool**
- **BUT for the applications we have in mind there are problems with existing (traditional) methods**
- **It is necessary with a new concept**

- **AIM: Classify forest land and/or wetland**
- **Overlapping feature distributions**
- **Non-normality**
- **Low accuracy**
- **Feasibility at local area**
- **Biased area estimation**

- **According to species composition forest land is defined into 32 classes**
 - **In-situ measurements**
 - **Multispectral and multitemporal satellite images give us a 12-dimensional feature vector**
 - **Spatial resolution= pixel size= 25 m**
-

- **Define the target function**
(in this case, probabilities of correct classification)
- **Denoise the images (wavelet transformations)**
- **Remove outliers from reference data**
- **Calculate the information values in the components in the feature vector (e.g. different bands)**
- **Determine a proper metric**
- **Determine prototypes for the classes**
- **Run a nonparametric classification so that the target function is maximized**
- **Declare the quality of classification result by using probability matrices**

End-user response

- **The probability matrices give quality statements on scene-level (180x180km)**
 - **Not sufficient for the end users**
 - **More local information is required**
-

- Calculate probabilities for classes at pixel level
- Calculate entropy for each pixel

Density estimation using i.i.d. X_1, X_2, \dots, X_n

Joint distribution $(X_i, Y_n(i)) \equiv (X_i, n \| B(X_i, R_n(i)) \|) \rightarrow (X, Y)$

Conditional distr. $Y | X = x \sim \exp[g(x)]$

Conditional expectation $E[Y_n(i) | X_i = x] \rightarrow \frac{1}{g(x)}$

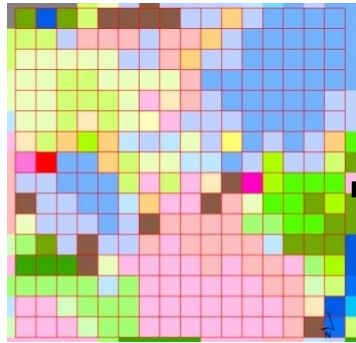
Density estimation

$$\hat{g}(x) = \frac{1}{n \| B(x, R_n(i)) \|} \propto \frac{1}{d^p(x, R_n(i))}$$

Density estimation of class k $\hat{f}_k(x) = \frac{1}{n_k \|B(x, R_n(k, x))\|}$, where $n = \sum n_k$

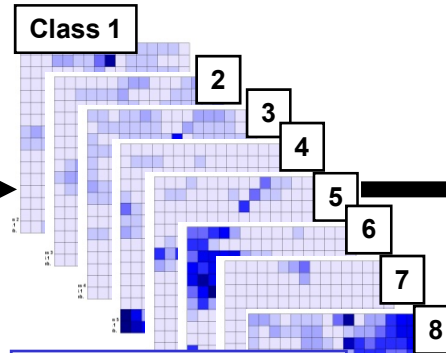
$$\begin{aligned}
 P(\text{pixel at } x \in \text{class } k) &= \frac{\pi_k \hat{f}_k(x)}{\sum_j \pi_j \hat{f}_j(x)} \\
 &= \frac{\pi_k}{\frac{n_k}{n} \|B(x, R_n(k, x))\|} \bigg/ \sum_j \frac{\pi_j}{\frac{n_j}{n} \|B(x, R_n(j, x))\|} \\
 \left(\frac{n_k}{n} \approx \pi_k\right) &= \frac{1}{d^p(x, R_n(k, x))} \bigg/ \sum_j \frac{1}{d^p(x, R_n(j, x))}
 \end{aligned}$$

The concept of probabilistic classifiers, example - a cost-efficient method for terrestrial monitoring

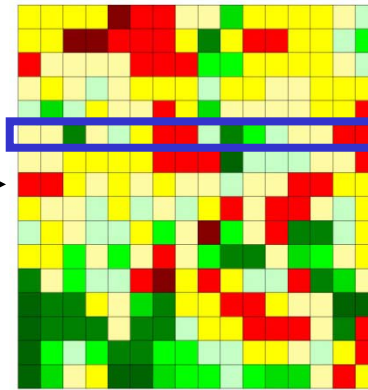


Preliminary classification

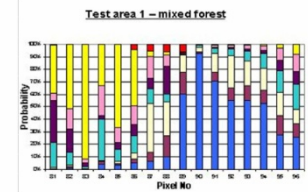
"mixed forest"
tree species & age
8 classes, pixel size
25 x 25 m



Calculation of pixelwise probability per class
($\Sigma = 1$)

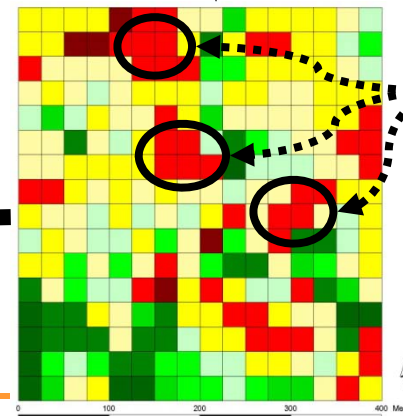


Calculation of entropy
red = classif. accuracy NOT OK
green - yellow = OK



Probability per class in pixel row

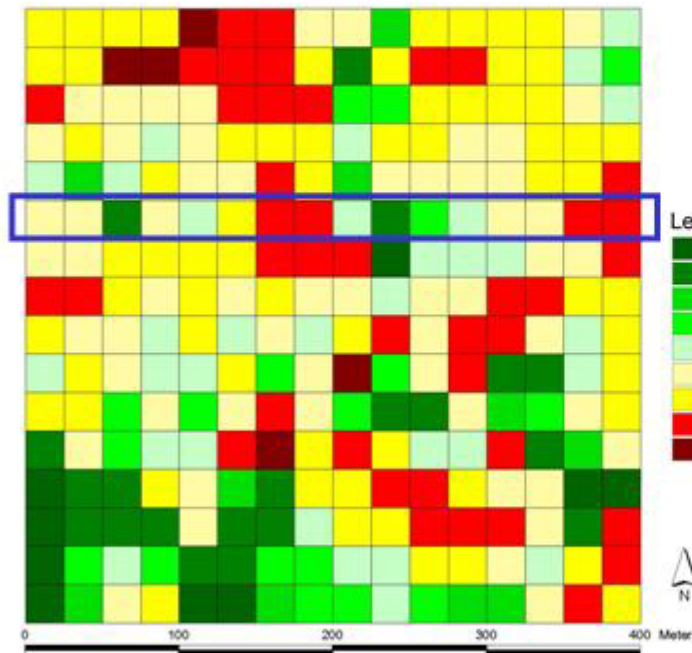
New classification with improved accuracy



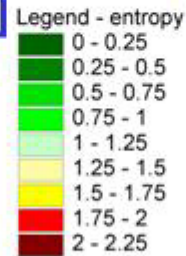
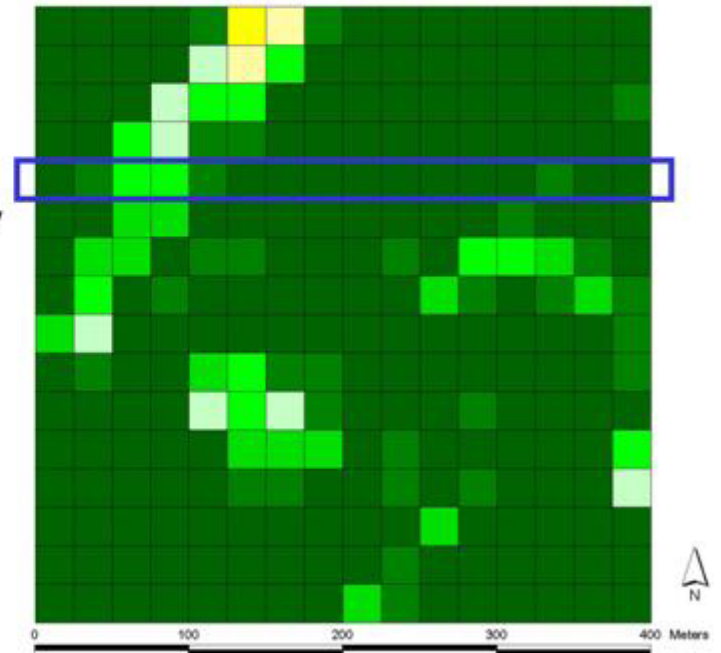
Additional field plots in selected areas

Entropy

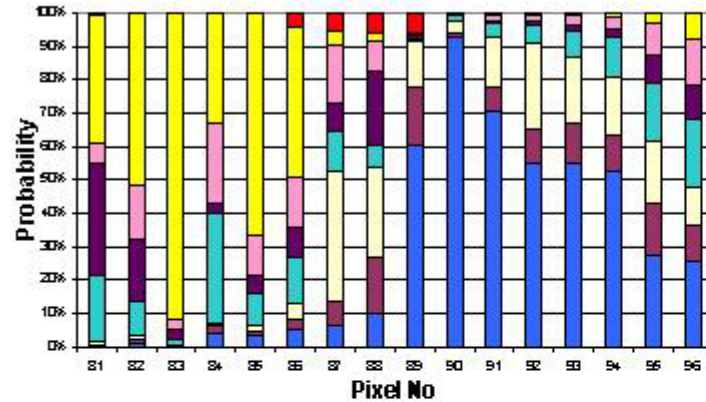
Test area 1 – mixed forest



Test area 2 – pine dominated forest

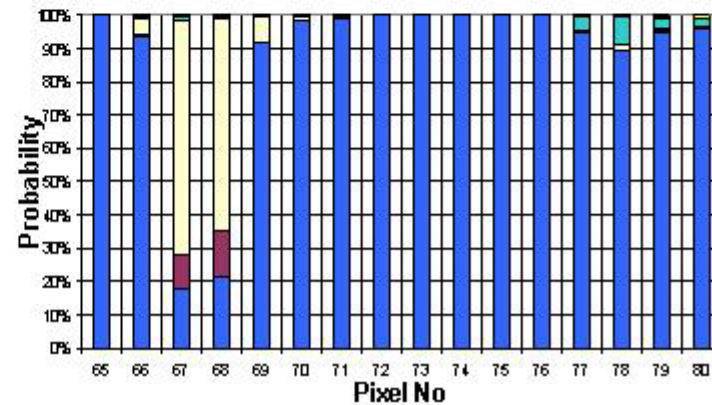










Test area 1 – mixed forest



**Pixel-wise probabilities in pixel row
($\Sigma = 1$)**

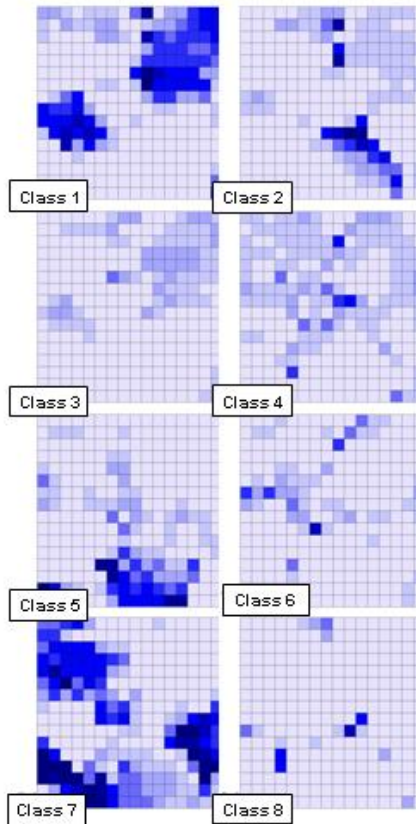
Test area 2 – pine dominated forest



Class	Colorscheme indigram
8	
7	
6	
5	
4	
3	
2	
1	

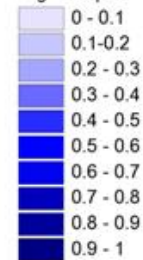
Probability Map

Test area 1 – mixed forest

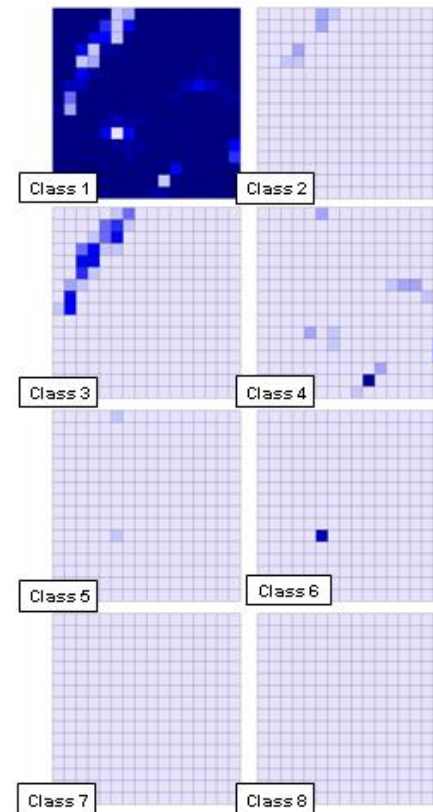


Pixel-wise probabilities
($\Sigma = 1$)

Legend - probability

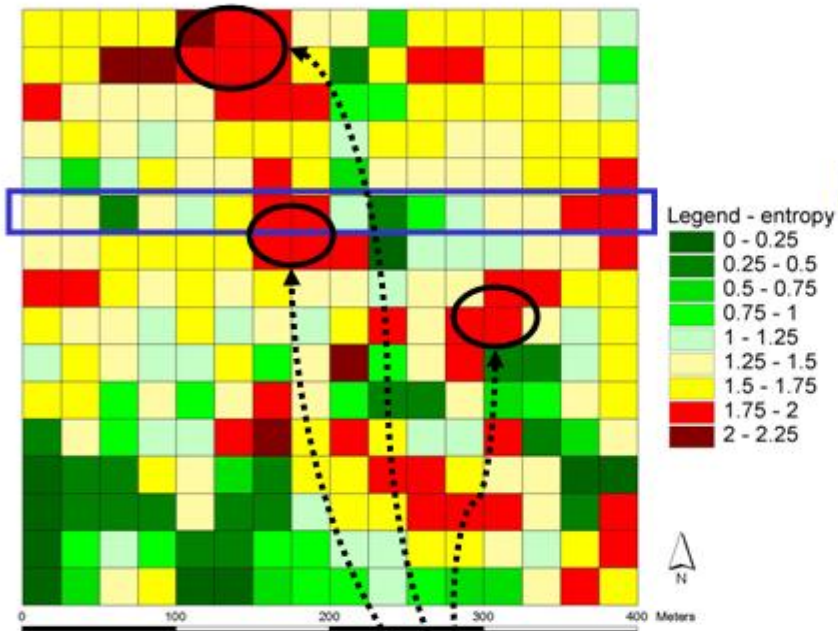


Test area 2 – pine dominated forest

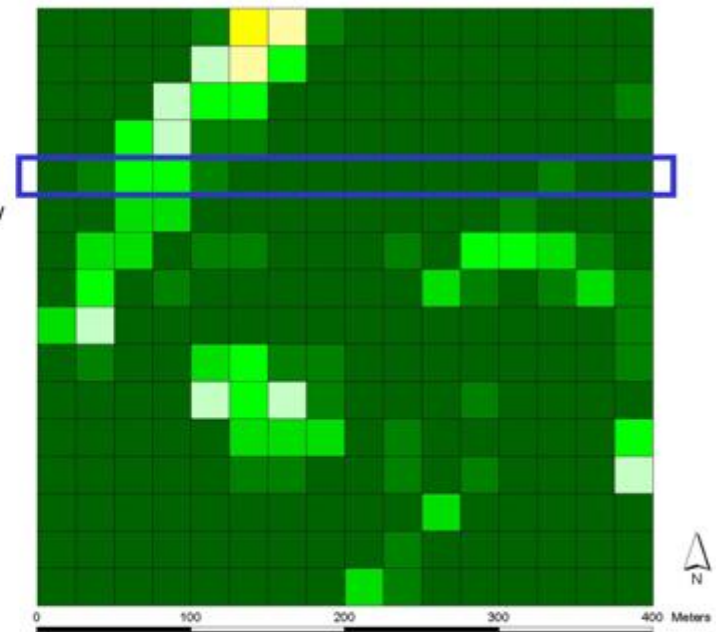


Entropy

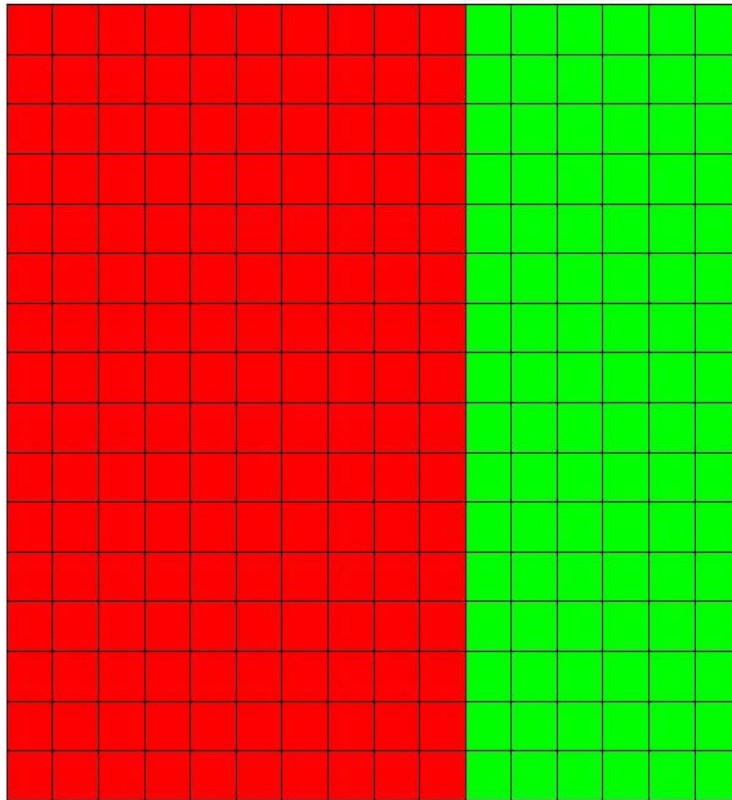
Test area 1 – mixed forest



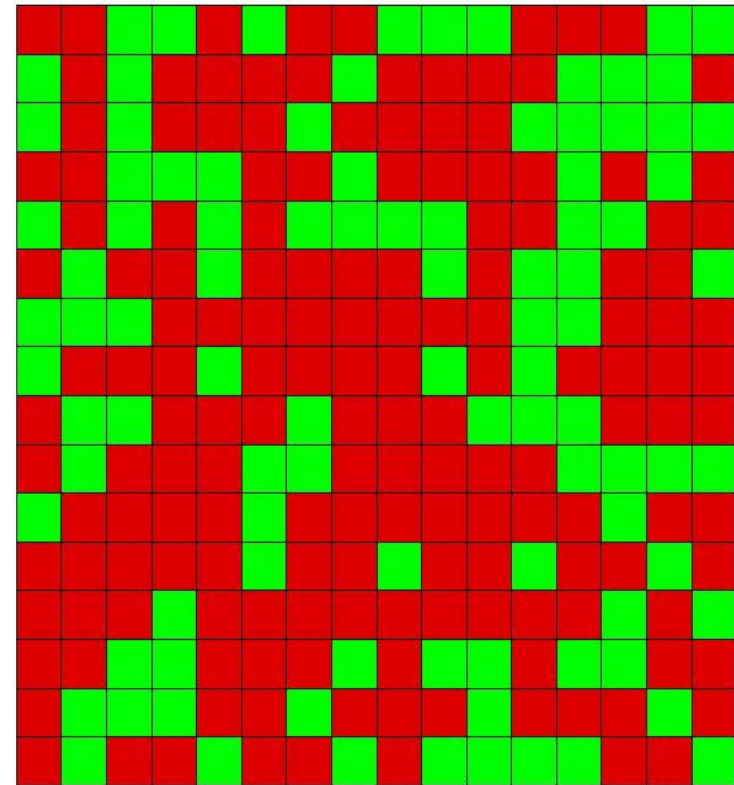
Test area 2 – pine dominated forest



Field investigations in selected areas



Trad. classification, Area est.=160
Red= prob.0.8 and green=prob.0.4
Prob.classifier, Area est.=134.4
Accuracy??????



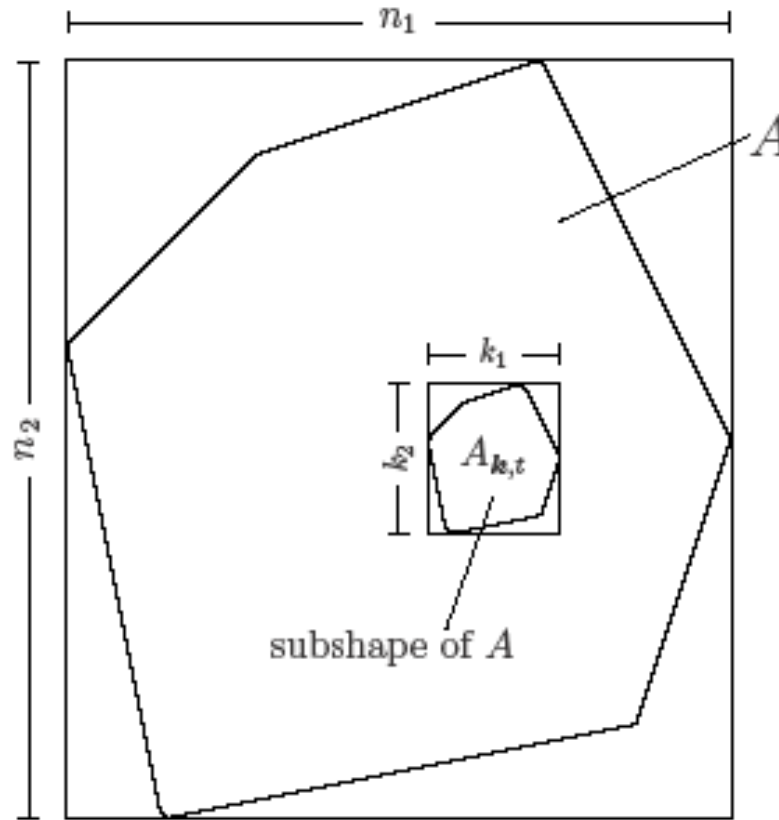
Trad. classification, Area est.=160
Red= prob.0.8 and green=prob. 0.4
Prob.classifier, Area est.=134.4
Accuracy??????

Area (both global and local), amount of dead wood,

Subsampling methods to estimate the variance of

- **sample means**
- **sample totals**
- **ratios of means or totals**

Consistent under stationarity and m -dependence



\bar{X}_A = mean of the process $\{X_{i,j}\}$ over the region A

\bar{X}_t = mean value over the subregion A_t

t_n = number of subregions

S = "size" of the region A , s = "size" of each subregion A_t

$$\bar{\bar{X}} = \frac{1}{t_n} \sum_{1}^{t_n} \bar{X}_t \quad \gamma_n = \text{Var}(\sqrt{S} \bar{X}_A)$$

$$\hat{\gamma}_n = \frac{S}{t_n} \sum_{1}^{t_n} \left(\bar{X}_t - \bar{\bar{X}} \right)^2 \quad \longrightarrow \quad \hat{\gamma}_n - \gamma_n \longrightarrow 0$$

Errors in forest and peatland masks

- **Within the peatland mask (defined by a topographic map) about 30% is in fact forest land**
 - **Within the forest mask (defined by the same topographic map) about 6-8% is in fact peatland**
 - **As a consequence performing the forest classification within the forest mask may give misleading results**
 - **Solution: try to make the classification of forest land and peatland simultaneously.**
-

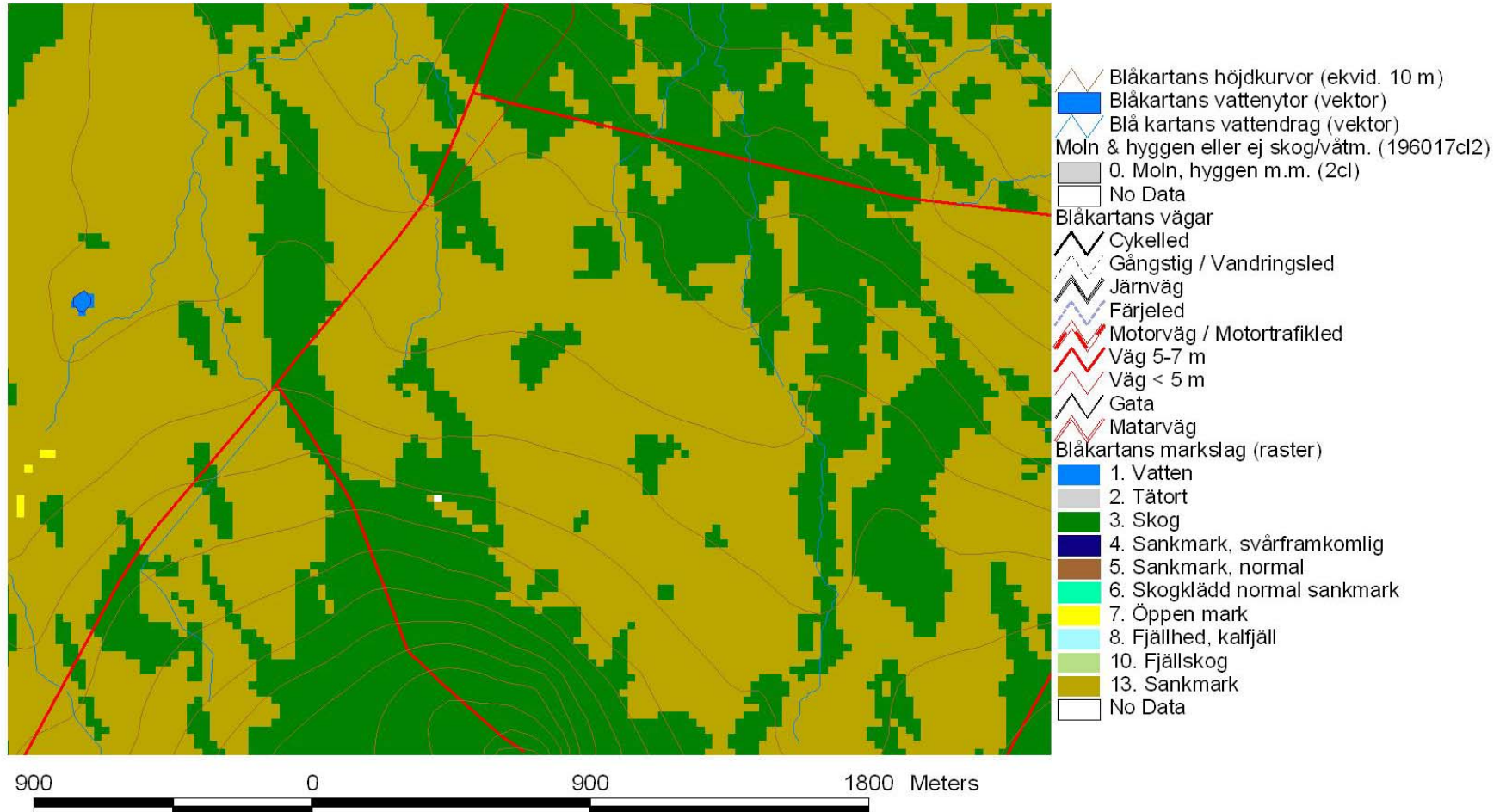
- ❑ Information from TOPO-index and neighboring pixels
 - ⇒ **A priori** probabilities for the classes at pixel level

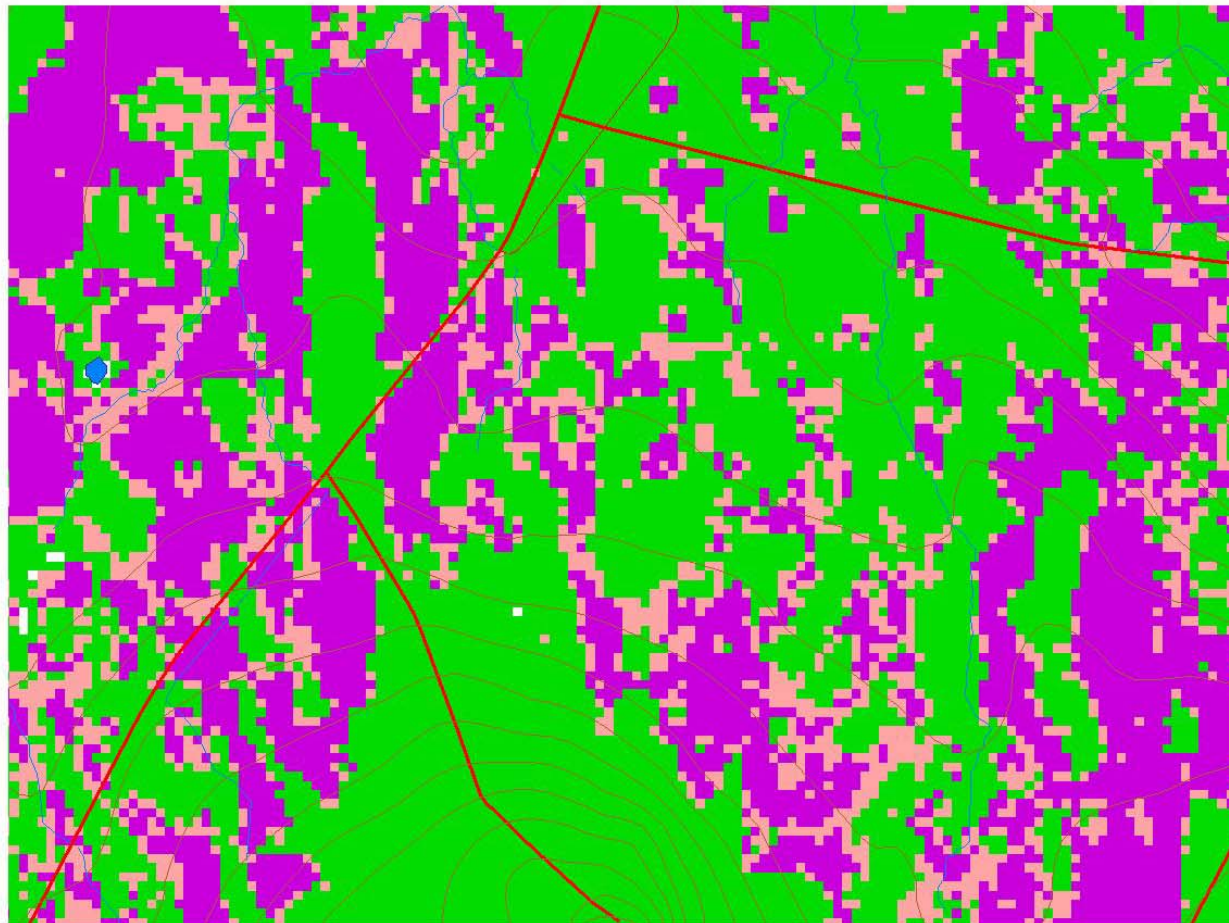
- ❑ A priori probabilities & previously calculated probabilities
 - ⇒ **A posteriori** probabilities for the classes at pixel level
 - ⇒ New "Probability Classifier"

- ❑ The classes are aggregated up to:
 - FOREST** and **PEATLAND**

- ❑ Calculate entropy and determine the thresholds so that the pixels belong to the following zones:
 - FOREST**, **PEATLAND**, or **Transition Zone**

- ❑ The probability vectors are recalculated with respect to the zones





-  Blåkartans höjdkurvor (ekvid. 10 m)
-  Blåkartans vattenytor (vektor)
-  Blå kartans vattendrag (vektor)
- Moln & hyggen eller ej skog/våtm. (196017cl2)
-  0. Moln, hyggen m.m. (2cl)
-  No Data
- Blåkartans vägar
-  Cykelled
-  Gångstig / Vandringsled
-  Järnväg
-  Färjeled
-  Motorväg / Motortrafikled
-  Väg 5-7 m
-  Väg < 5 m
-  Gata
-  Matarväg
- Klassindelning 4 prob.skog (Prob_sk19617)
-  0 - 0.3
-  0.3 - 0.7
-  0.7 - 1
-  No Data

900 0 900 1800 Meters

- **Estimation/classification of changes**
- **Cost-efficient monitoring systems**
- **Identification of sparse events/ hot spot detection**
- **Source apportionment models (nitrogen-phosphorus leakage)**
- **Climate change: scenarios based on changes in e.g. species composition**