ROC analysis, a method to assess binary decision rules

David Makowski
INRA

Outline

1. What is a binary decision rule?

2. ROC analysis, a method to assess the accuracy of binary decision rules

3. An example: assessment of decision rules for the control of sclerotinia

4. Exercice with R:

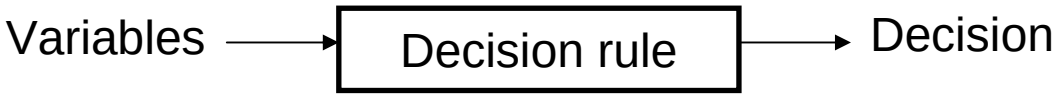
   Assessment of models for categorizing soft wheat fields according to their grain protein content

1. What is a binary decision rule?

1. What is a binary decision rule?

Decision rule

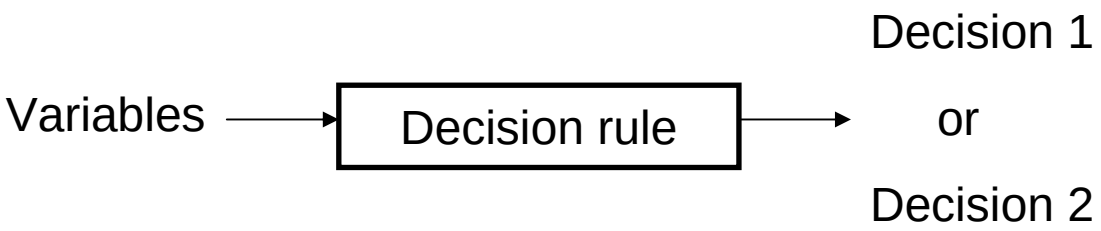
« A rule for taking decisions in function of some variables ».



1. What is a binary decision rule?

Binary decision rule

« A rule for choosing among two decisions ».



1. What is a binary decision rule?

Examples of binary decision rules

- **Apply a chemical treatment / No chemical treatment**
- **Sow cultivar 1 / Sow cultivar 2**
- **Apply fertilizer / No application**

…

1. What is a binary decision rule?

Binary decision rule based on an indicator and a decision threshold

« I apply fungicide if the indicator is higher than a decision threshold »

« I apply fungicide if *I* ≥ *S*. No application otherwise »

- Indicator ***I*** = measure or model prediction (ex: % diseased organs).
- Threshold ***S*** = Numerical value (ex: 20%).

1. What is a binary decision rule?

Optimization of

« I apply fungicide if *I* ≥ *S*. No application otherwise »

Two practical problems:

- Choose the best threshold *S* for a given indicator ***I***.
- Choose the best indicator among several candidates.

1. What is a binary decision rule?

A framework for assessing binary decision rules

1. Define a series of indicators (measured variables and/or models).

2. Define the range of variation for the threshold *S* associated to each indicator (*e.g* 0-100 % of diseased flowers).

3. Define one or several criteria for assessing the decision rules (*i.e* the combinations of all possible *I* and *S*).

4. Estimate the values of the criteria for each rule.

5. Choose the « best » rule.

2. ROC analysis

ROC = Receiver Operating Characteristic

ROC analysis

Notations

Y : a random variable taking the value 0 or 1 for a negative and positive response respectively.

I : a variable corresponding to the output of a given indicator.

S : a decision threshold.

Examples for *Y*

Y = 0 if the yield loss due to the disease is small, *Y*=1 otherwise.

Y = 0 if the percentage of diseased plants at harvest < 10%, *Y*=1 otherwise.

Y = 0 if weed biomass < 0.15 t/ha, *Y*=1 otherwise.

ROC analysis

n plots with *Y*=0 (*e.g*. % diseased plants at harvest < 10%).

m plots with *Y*=1 (*e.g.* % diseased plants at harvest ≥ 10%).

(i). Determine the value of the indicator *I* for each plot.

(ii). Define a decision threshold S.

(iii). **Sensitivity** = *Prob*(*I* ≥ S | *Y=1*) = 1 – **False negative rate**

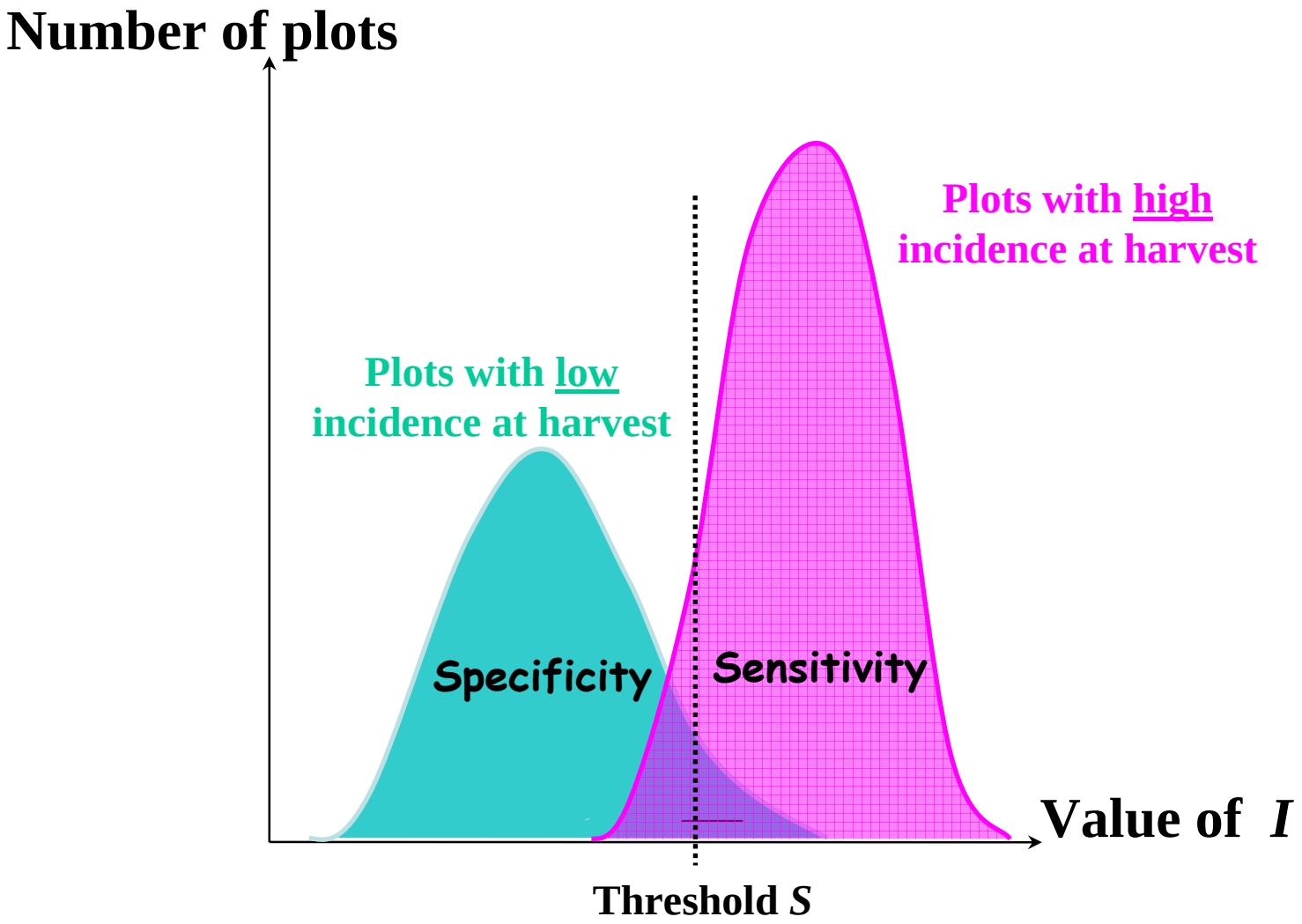
(iv). **Specificity** = *Prob*(*I* < S | *Y=0*) = 1 – **False positive rate**

(v). **ROC curve**: Sensitivity (S) *versus* 1 – Specificity (S)

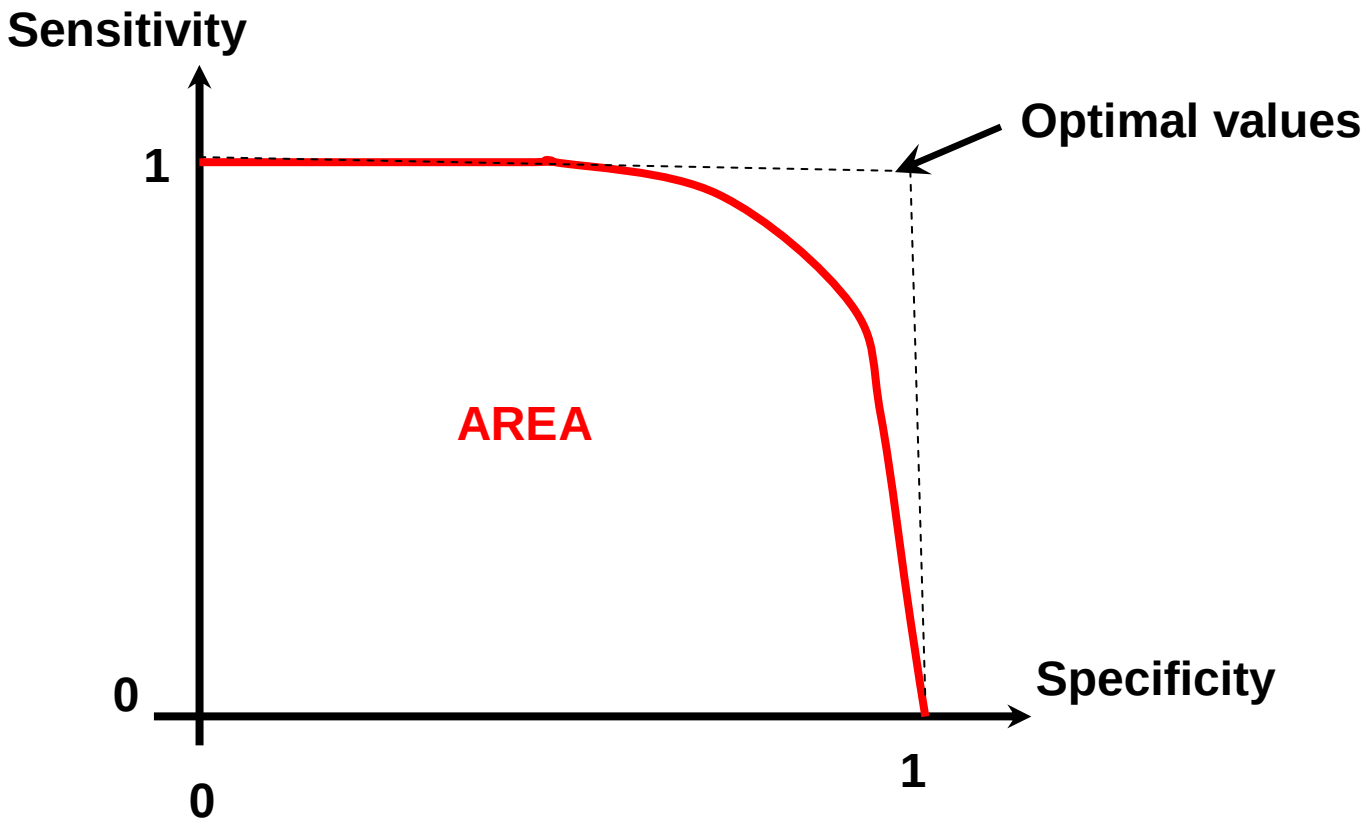
(vi). Estimate the area under the ROC curve (**AUC**) for each indicator *I*.

If AUC~0.5, the indicator is not useful (not better than random decisions).

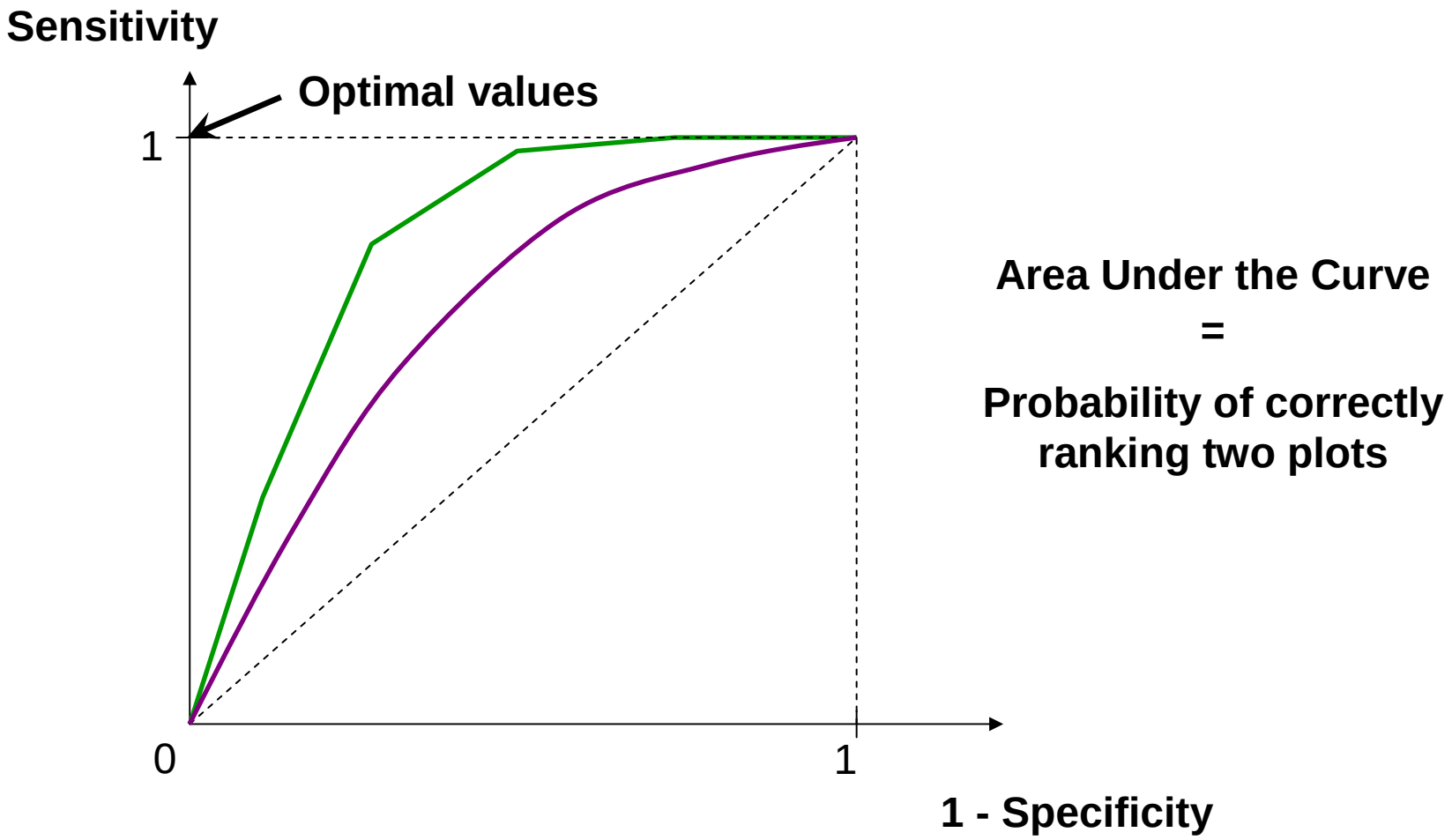
ROC analysis



2. ROC analysis



2. ROC analysis



3. An example: assessment of decision rules for the control of sclerotinia

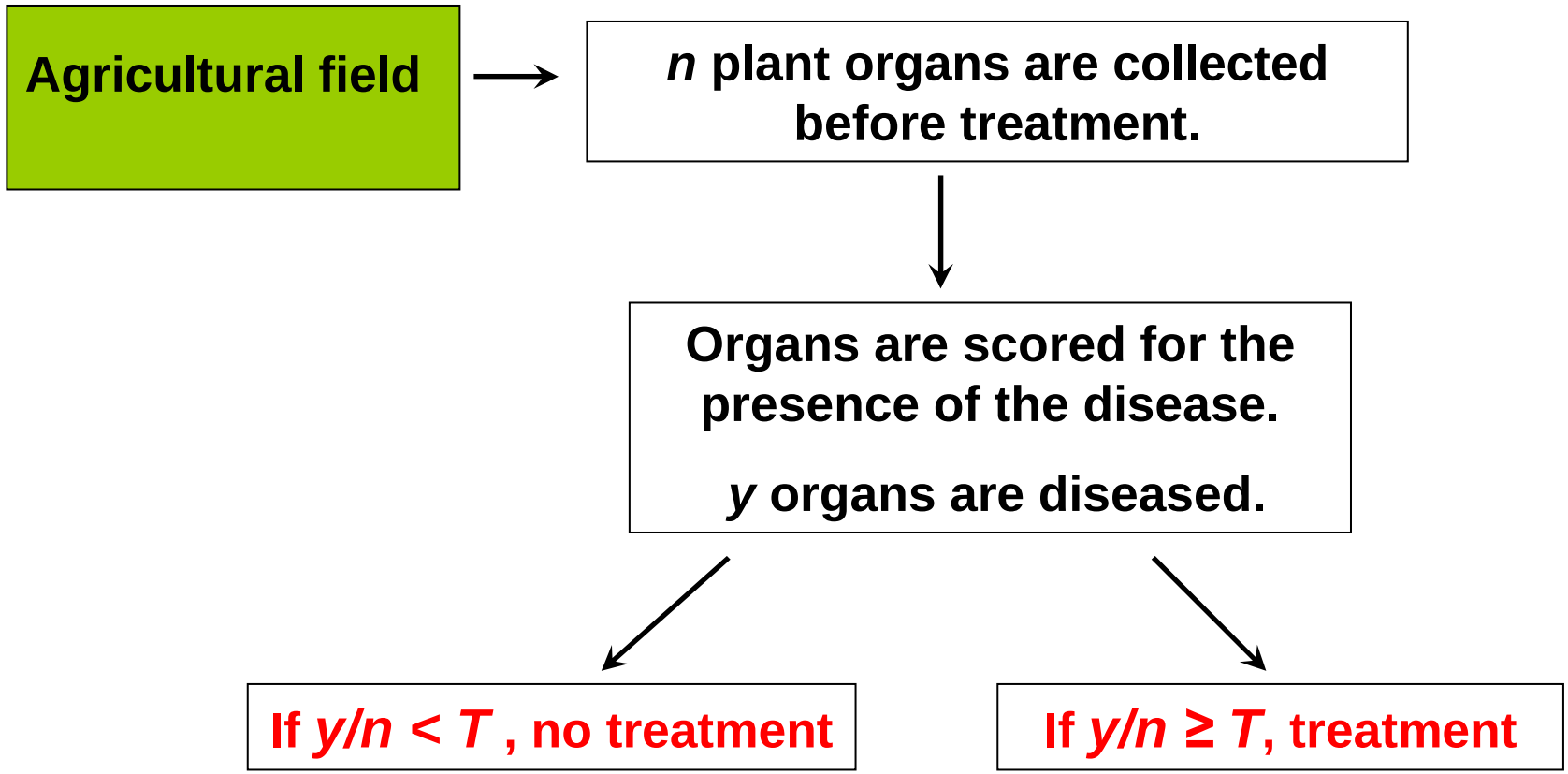
3. An example

Sclerotinia sclerotiorum, Lib., de Bary, in oilseed rape crops

- **High variability of disease incidence across sites and years.**
- **High yield losses if disease incidence at harvest > 10%.**
- **Efficient chemical treatments exist, but are not always required.**

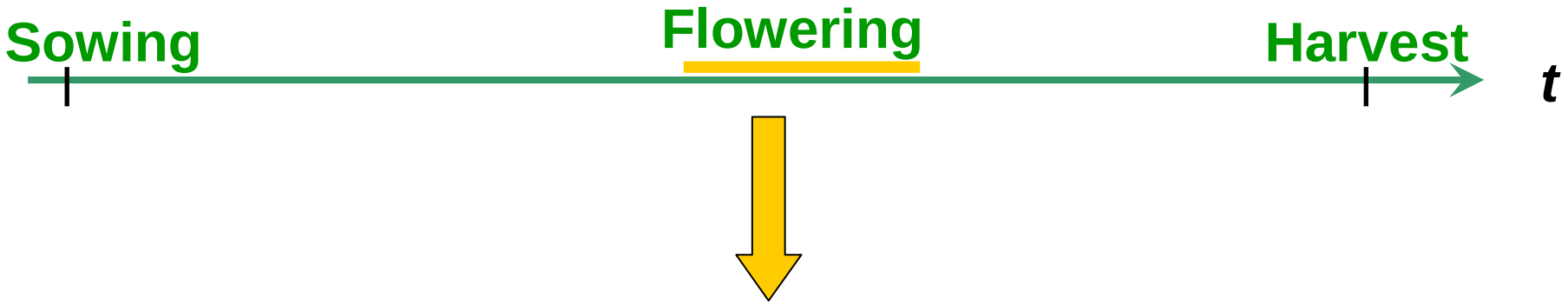
3. An example

Rule 1. Indicator I_1 = measured proportion of diseased plant organs



3. An example

In this example, organs = flowers



***n* collected flowers**

↓

Incubation in Petri dishes

↓

***y* diseased flowers**

If $y / n \geq T$ treatment

else no treatment

3. An example

Rule 2. Indicator I_2 = sum of risk points

Risk factor	Level	Points
Number of oil-seed crops during the last ten years	>5	30
	3-5	20
	2-3	10
	1	0
Other host crops during the last five years	Yes	15
	No	0
Level of infection in the last crop	High	15
	Moderate	5
	Low	0
Type of field	Wet	10
	Dry	0
Plant density	High	10
	Normal	5
	Low	0
Rain in the last month before flowering	More than normal	10
	Normal (50-60 mm)	5
	Less than normal	0

3. An example

Rule 3. Indicator I_3 = output of a logistic model

% diseased flowers at flowering

Sum of risk points at flowering

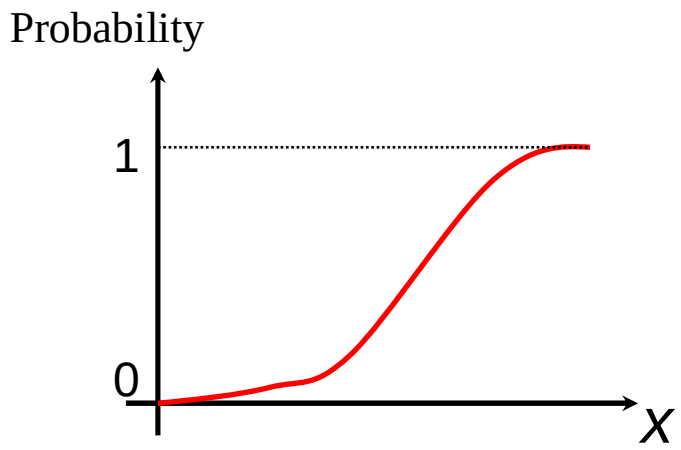
MODEL

Probability that the disease incidence at harvest is higher than 10%

Rule 3. Indicator I_3 = output of a logistic model

Logistic model

$$z = \frac{\exp\left(\theta_0 + \theta_1 x_1 + \theta_2 x_2\right)}{1 + \exp\left(\theta_0 + \theta_1 x_1 + \theta_2 x_2\right)}$$



Three rules for deciding about a chemical treatment at flowering

If $I_1 \geq S$, a treatment is recommended

If $I_2 \geq S$, a treatment is recommended

If $I_3 \geq S$, a treatment is recommended

Which rule is the best?

Two types of error

Type 1. False positive rate = 1 - Specificity

I* ≥ *S (a treatment was recommended)

but % diseased plants at harvest < 10%

(a treatment was not required)

Type 2. False negative rate = 1 – Sensitivity

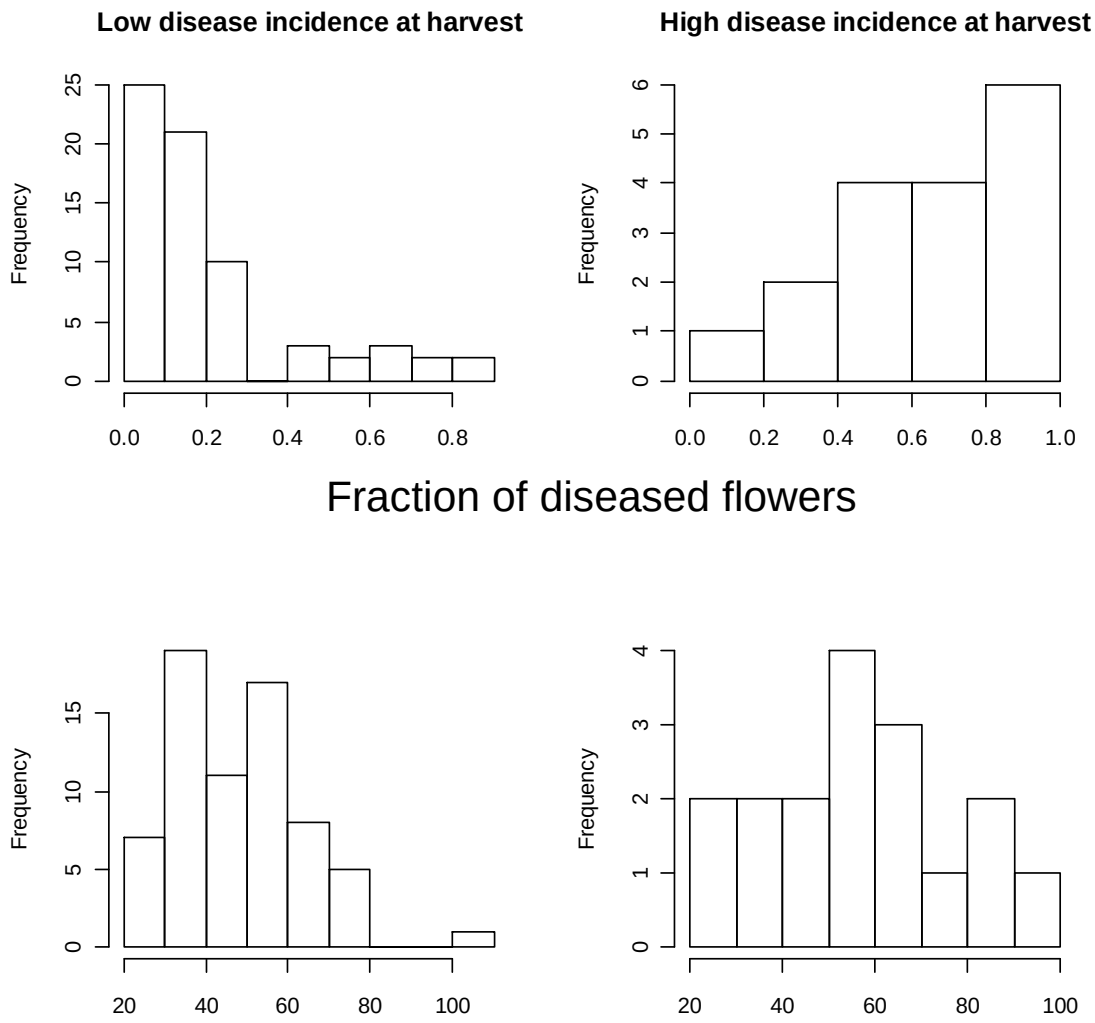
I* < *S (a treatment was not recommended)

but % diseased plants at harvest ≥ 10%

(a treatment was required)

3. An example

Data from 85 experimental plots in France



R code for fitting the logistic model

```
TAB<-read.table("f:\\ Exemples\\Sclero0203.txt",header=T,sep="\t")
TAB<-TAB[is.na(TAB[,1])==F,]
Ind.1<-TAB$KIT
Ind.2<-TAB[,3]+TAB[,4]+TAB[,5]+TAB[,6]+TAB[,7]+TAB[,8]+TAB[,9]+TAB[,10]+TAB[,11]+TAB[,12]
Incidence.t<-0.10
Incidence<-TAB$TxAttNT
Incidence[Incidence<Incidence.t]<-0
Incidence[Incidence>=Incidence.t]<-1
Fit<-glm(Incidence~Ind.1+Ind.2,family=binomial)
```

glm = R function for fitting generalized linear models (e.g logistic, Poisson)

3. An example

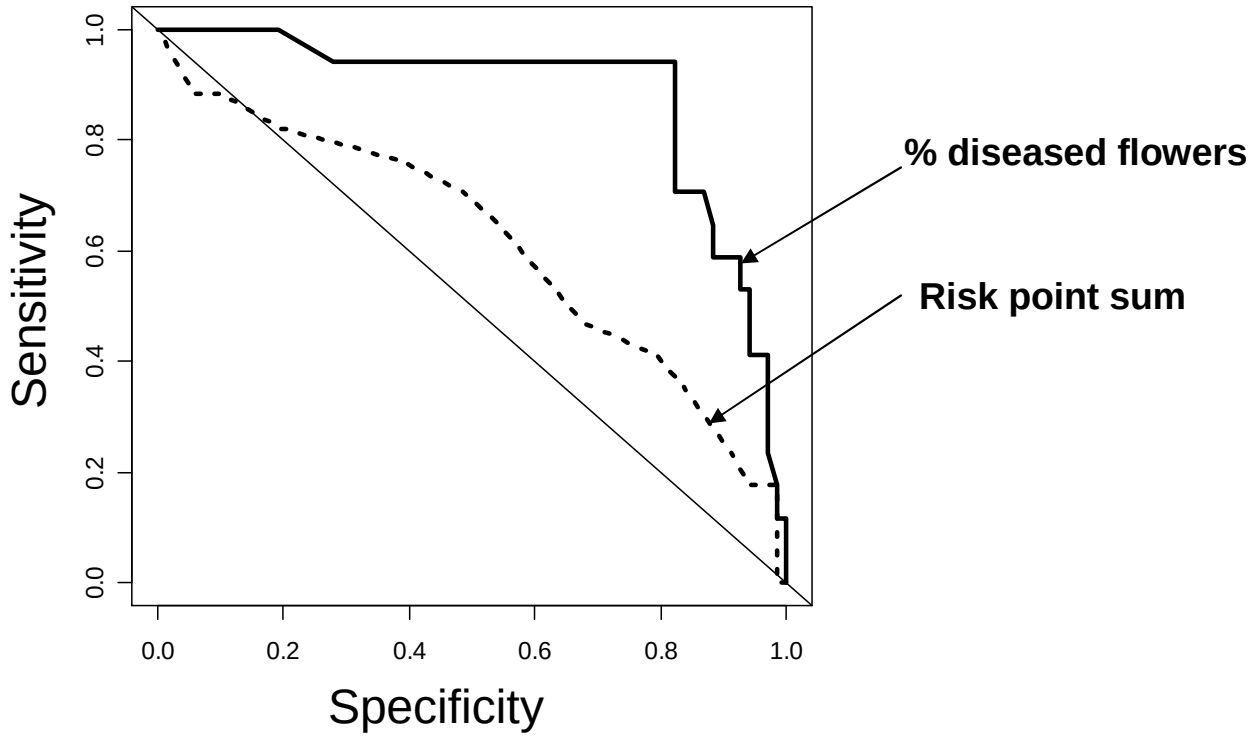
```
> print(summary(Fit))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.31581    1.20336  -3.586 0.000335 ***
Ind.1        5.27346    1.21518   4.340 1.43e-05 ***
Ind.2        0.01356    0.01800   0.753 0.451329
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R code for ROC analysis for rule 1

```
library(ROCR)
pred<-prediction(Ind.1,Incidence)
perf<-performance(pred,"sens","spec")
spec.1<-perf@"x.values"[[1]]
sens.1<-perf@"y.values"[[1]]
plot(spec.1,sens.1, ylab="Sensibilité", xlab="Spécificité", type="l",lty=1,lwd=3)
abline(1,-1)
```

3. An example

ROC curves for rules 1 and 2

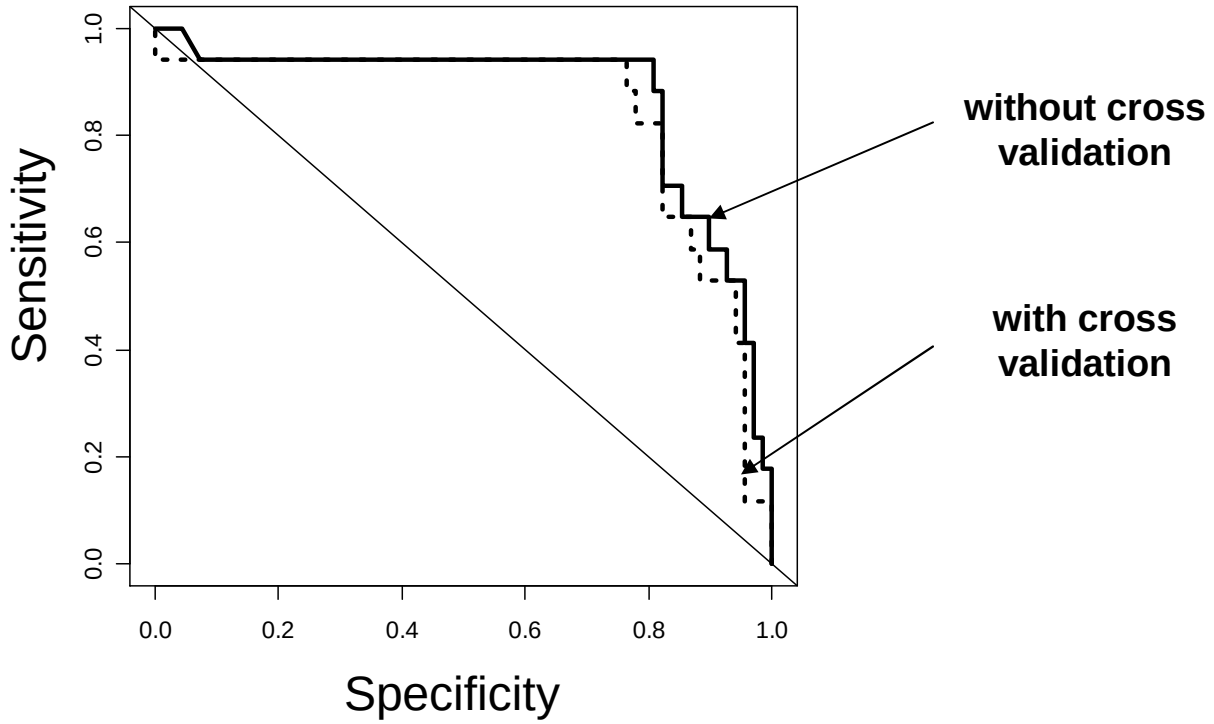


R code for ROC analysis with cross validation

```
Pred.cv<-NA
for (i in (1:length(TAB[,1]))) {
TAB.est.i<-data.frame(Ind.1[-i],Ind.2[-i],Incidence[-i])
TAB.pred.i<-c(Ind.1[i],Ind.2[i])
Fit.cv<-glm(TAB.est.i$Incidence~TAB.est.i[,1]+TAB.est.i[,2],family=binomial,data=TAB.est.i)
Para<-as.vector(Fit.cv$coefficients)
Pred.i<-exp(Para[1]+Para[2]*TAB.pred.i[1]+Para[3]*TAB.pred.i[2])/(1+
exp(Para[1]+Para[2]*TAB.pred.i[1]+Para[3]*TAB.pred.i[2]))
Pred.cv<-c(Pred.cv, Pred.i)
}
pred<-prediction(Pred.cv[-1],Incidence)
perf<-performance(pred,"sens","spec")
```

3. An example

ROC curves for rule 3



Area under the ROC curves

Indicator	AUC
% diseased flowers	0.88
Point sum	0.62
Logistic (without cross validation)	0.87
Logistic (with cross validation)	0.85

5. Exercice with R:

Assessment of models for categorizing soft wheat fields according to their grain protein content

Four candidate indicators:

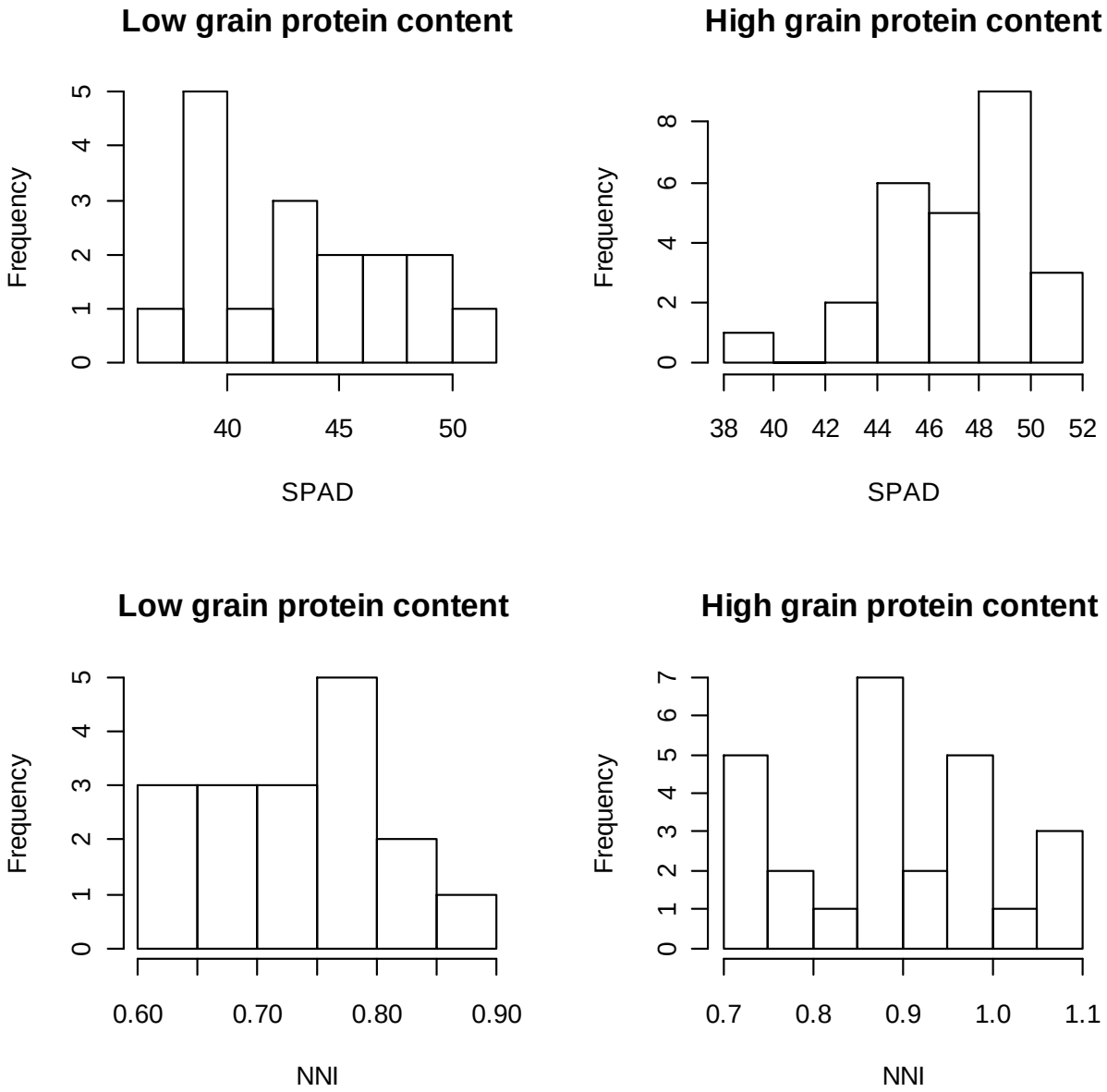
i. **Transmitance**
ii. **Nitrogen nutrition index**
iii. **Model 1 (dynamic crop model)**
iv. **Model 2 (static crop model including two input variables)**

Objective: Identify plots with high grain protein content (>11.5%)

Which indicator is the best? What is its optimal decision threshold?


600

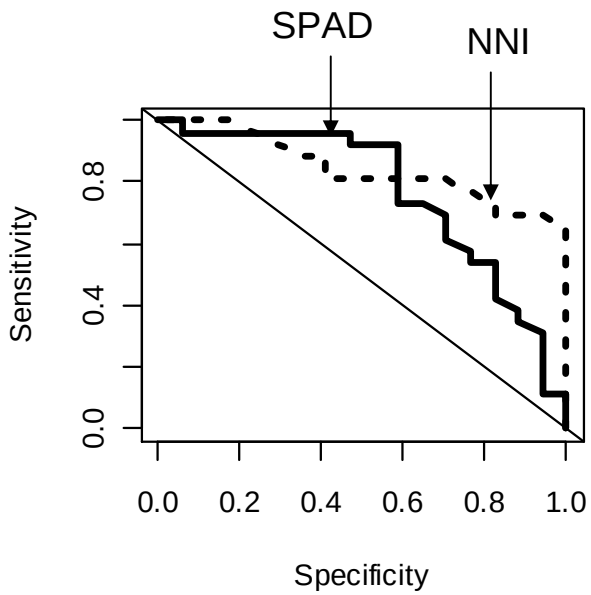
```
library(ROCR)

#Read an external data file
TAB<-read.table("f:\\David\\Enseignements\\FormationPologne\\dataAgralys.txt",header=T,sep="\t")
print(TAB)

#Grain protein threshold
GPC.t<-11.5

#Variable of reference (binary variable Y)
GPC<-TAB$Protein
GPC[GPC<GPC.t]<-0
GPC[GPC>=GPC.t]<-1

#Indicators
Ind.1<-TAB$SPAD
Ind.2<-TAB$NNI
Ind.3<-TAB$Model_1
Ind.4<-TAB$Model_2

#Some graphs
par(mfrow=c(2,2))
hist(Ind.1[GPC==0], xlab="SPAD", main="Low grain protein content")
hist(Ind.1[GPC==1], xlab="SPAD", main="High grain protein content")
hist(Ind.2[GPC==0], xlab="NNI", main="Low grain protein content ")
hist(Ind.2[GPC==1], xlab="NNI", main="High grain protein content ")
```

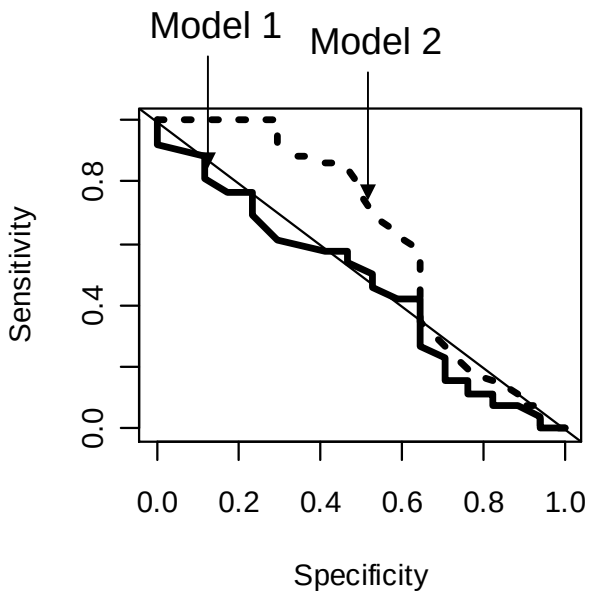

Low grain protein content
Frequency
0
1
2
3
4
5
40
45
50
SPAD
High grain protein content
Frequency
0
2
4
6
8
38
40
42
44
46
48
50
52
SPAD
Low grain protein content
Frequency
0
1
2
3
4
5
0.60
0.70
0.80
0.90
NNI
High grain protein content
Frequency
0
1
2
3
4
5
6
7
0.7
0.8
0.9
1.0
1.1
NNI

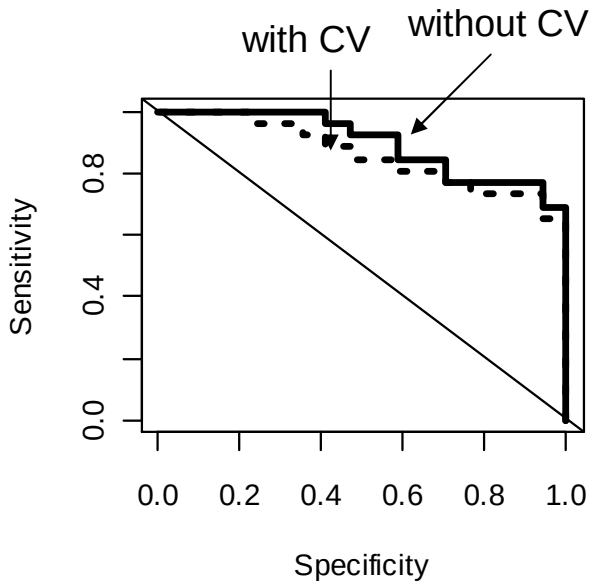
```
####ROC analysis for Ind.1####

pred<-prediction(Ind.1,GPC)
perf<-performance(pred,"auc")

#Area under the ROC curve
auc.1<-perf@"y.values"
print("AUC for Indicator 1")
print(auc.1)

#Sensitivity and specificity
perf<-performance(pred,"sens","spec")
print(perf)
spec.1<-perf@"x.values"[[1]]
sens.1<-perf@"y.values"[[1]]

#ROC curve
plot(spec.1,sens.1, ylab="Sensitivity", xlab="Specificity", type="l",lty=1,lwd=3)
abline(1,-1)

#Threshold
print(perf@"alpha.values"[[1]][spec.1>0.65 & sens.1>0.65])
```

```
#Logistic regressions
#Combination of Ind.1 and Ind.2
Fit<-glm(GPC~Ind.1+Ind.2,family=binomial)

print("Combination of Ind.1 and Ind.2")
print(summary(Fit))

print("ROC analysis for the combinations of indicators without cross-
validation")

#ROC analysis for Ind.1 + Ind.2

pred<-prediction(Fit$fitted.values,GPC)
perf<-performance(pred,"auc")
auc<-perf@"y.values"
print("AUC for Ind.1 + Ind.2")
print(auc)
perf<-performance(pred,"sens","spec")
spec.1<-perf@"x.values"[[1]]
sens.1<-perf@"y.values"[[1]]
plot(spec.1,sens.1, ylab="Sensitivity", xlab="Specificity", type="l",lty=1,lwd=3)
abline(1,-1)
```

```
print("ROC analysis for the combinations of indicators with cross-validation")

#Initialization

Pred.cv<-NA

for (i in (1:length(TAB[,1]))) {

#New tables
TAB.est.i<-data.frame(Ind.1[-i],Ind.2[-i],GPC[-i])
TAB.pred.i<-c(Ind.1[i],Ind.2[i])

#Combination of Ind.1 and Ind.2
Fit.cv<-glm(TAB.est.i$GPC~TAB.est.i[,1]+TAB.est.i[,2],family=binomial,data=TAB.est.i)
Para<-as.vector(Fit.cv$coefficients)
Pred.i<-
exp(Para[1]+Para[2]*TAB.pred.i[1]+Para[3]*TAB.pred.i[2])/(1+exp(Para[1]+Para[2]*TAB.pr
ed.i[1]+Para[3]*TAB.pred.i[2]))
Pred.cv<-c(Pred.cv, Pred.i)

}
```







Indicator	AUC
SPAD	0.77
NNI	0.84
Model 1	0.46
Model 2	0.62
SPAD + NNI	0.86 (0.90)

References

Barbottin A, Makowski D, Le Bail M, Jeuffroy M-H, Bouchard Ch, Barrier C. 2008. Comparison of models and indicators for categorizing soft wheat fields according to their grain protein contents. *European Journal of Agronomy* 29, 159-183.

Hughes, G., McRoberts, N., Burnett, F.J., 1999. Decision-making and diagnosis in disease management. *Plant Pathol*. 48, 147-153.

Makowski D., Denis J-B., Ruck L., Penaud A. 2008. A Bayesian approach to assess the accuracy of a diagnostic test based on plant disease measurement. *Crop Protection* 27:1187-1193.

Makowski, D., M. Taverne, J. Bolomier, M. Ducarne. 2005. Comparison of risk indicators for sclerotinia control in oilseed rape. *Crop Protection* 24:527-531.

Pepe MS. 2003. The statistical evaluation of medical tests for classification and prediction. Oxford Statistical Science Series 28.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940-3941.

Swets, J.A., Dawes, R.M., Monahan, J., 2000. Better decisions through science. Scientific American 283, 70-75.