A spatial classification : application to ecological data

L. Bel[1], A. Bar-Hen[3], D. Allard[4] J-M., Laurent[2], R. Cheddadi[2]

[1]Probabilités, Statistique et Modélisation, Université Paris-Sud, Orsay, France

[2]Institut des Sciences de l'Evolution, CNRS and Université Montpellier II, France

[3]Institut National d'Agronomie, Paris, France

[4]Institut National de la Recherche Agronomique, Unité de Biométrie, Avignon, France

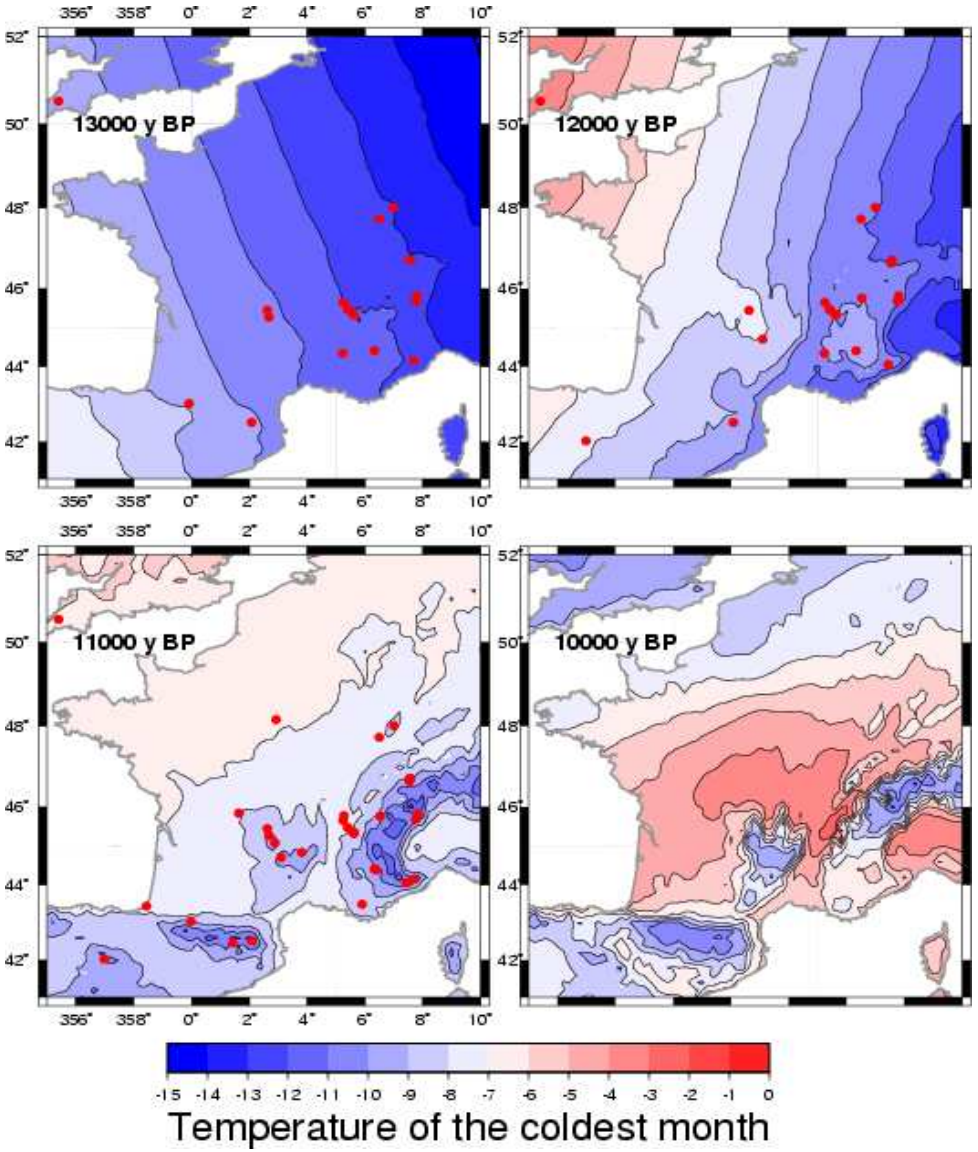
Paleoecological Questions

Paleoecology : reconstructing past environments and their evolution.
Pollen frequencies reproduce plant ranges.

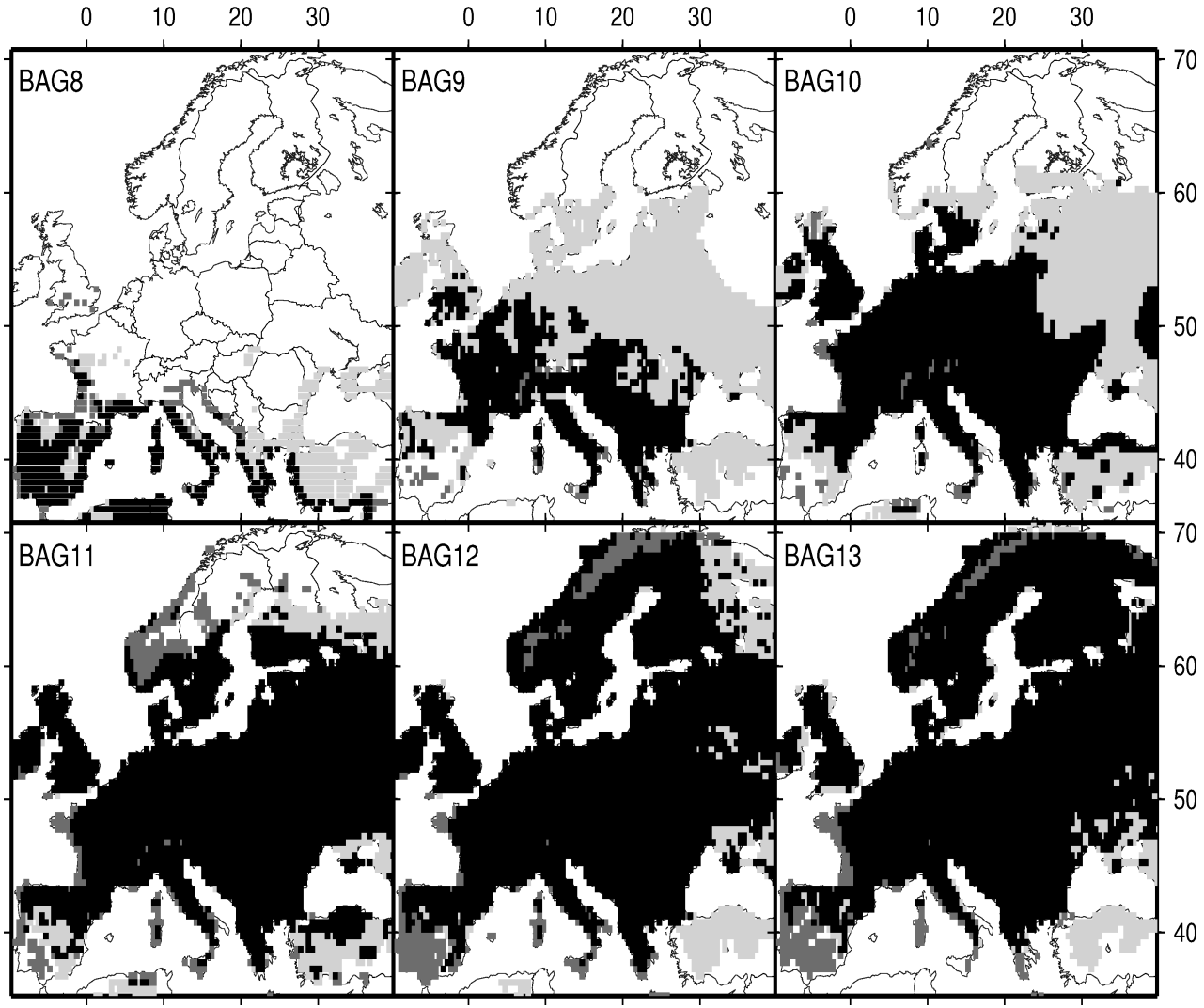
Goal : discriminate between absence, presence and abundance
of plant using pollen frequencies.

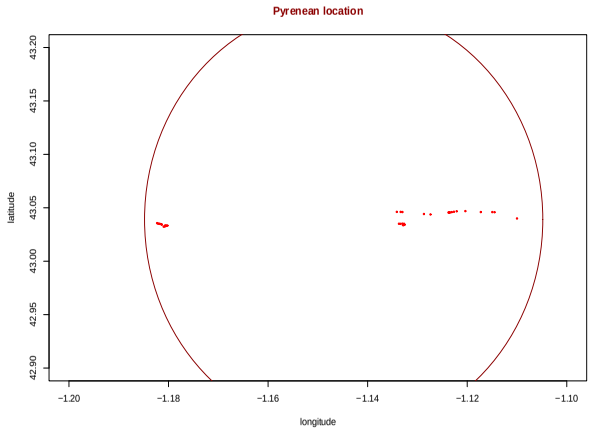
Supervised Classification with actual data as training set.
Sampling spatial scheme of pollen data very irregular.

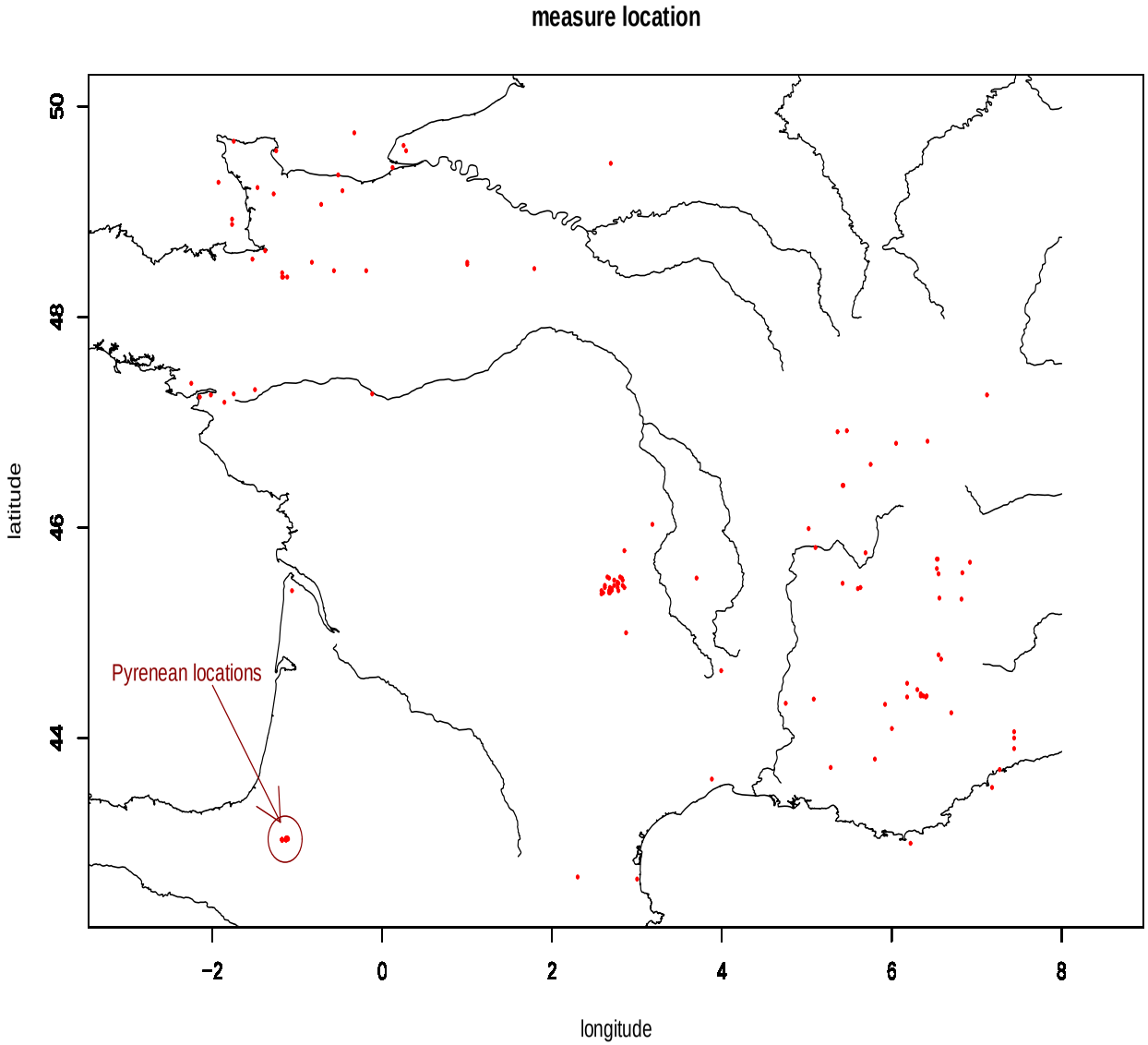
Example of climate reconstruction



European distribution of BAGs




Pyrenean location
latitude
43.20
43.15
43.10
43.05
43.00
42.95
42.90
−1.20
−1.18
−1.16
−1.14
−1.12
−1.10
longitude


measure location
latitude
50
48
46
44
−2
0
2
4
6
8
longitude
Pyrenean locations

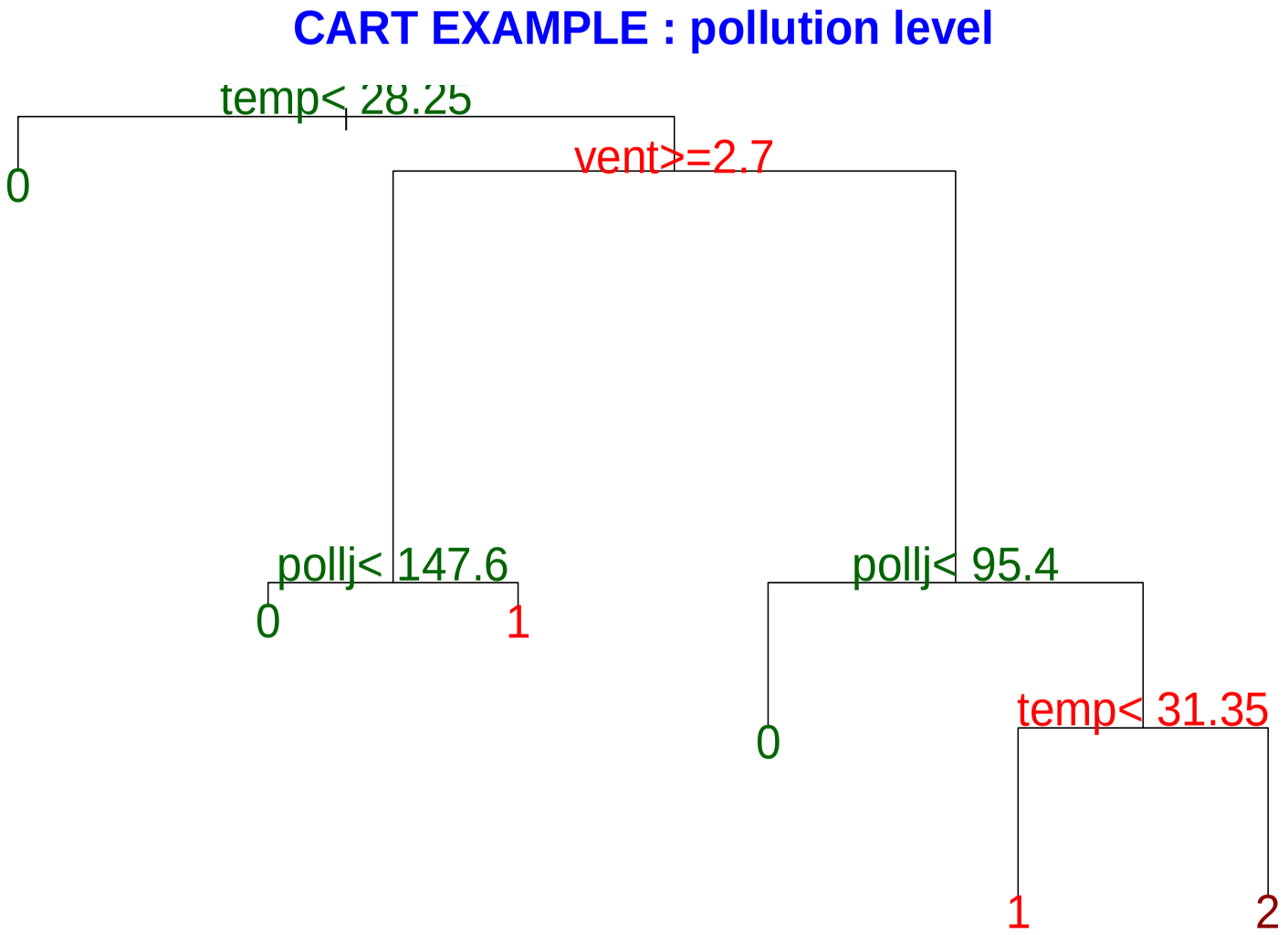
CART

Classification And Regression Tree

$X^1, \ldots X^p$: explanatory variables, Y: categorical explained variable
Binary tree through binary recursive partitioning

Example: predict Ozone pollution level from climatic variables

3 variables :
temperature (temp)
wind (vent)
O_3 previous day (pollj)
3 classes :
0 : low pollution
1 : medium pollution
2 : high pollution


CART EXAMPLE : pollution level
temp< 28.25
0
vent>=2.7
pollj< 147.6
0
1
pollj< 95.4
0
temp< 31.35
1
2

CART : algorithm

- Heterogeneity criterion

$$\text{Gini Index} : D_l = \sum_{i,j} p_{il} p_{jl} = 1 - \sum_i p_{il}^2$$

- Splitting rule

$$D_l - p_{l_+} D_{l_+} - p_{l_-} D_{l_-} > 0$$

- Criterion error prediction $R(T) = \mathsf{P}[T(X) \neq Y]$

- Penalized criterion $S(T) = \text{argmin}\{\widehat{R}(T) + \mu.\text{size}(T)\}$

Weighting CART

- Classically assume independence of the data

$$p_{il} = \frac{n_{il}}{n_{+l}} \qquad \widehat{R}(T) = \frac{1}{n} \sum_{\alpha=1}^{n} \mathbb{I}(T(X_\alpha) \neq Y_\alpha)$$

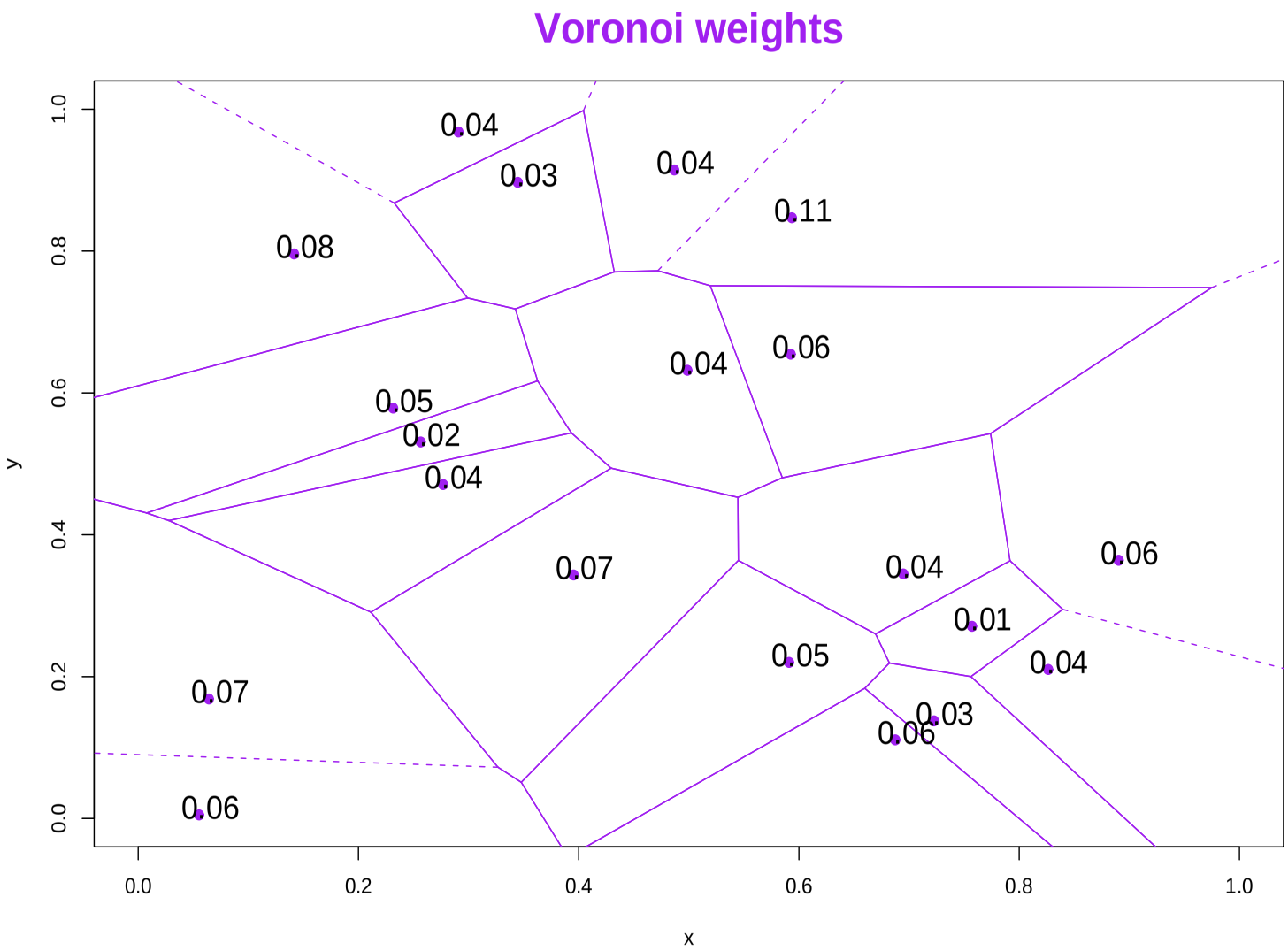
- Dependence through weights for individuals

$$p_{il} = \sum_{\alpha \in l} w_\alpha \mathbb{I}(Y_\alpha = i) \qquad \widehat{R}(T) = \sum_{\alpha=1}^{n} w_\alpha \mathbb{I}(T(X_\alpha) \neq Y_\alpha)$$

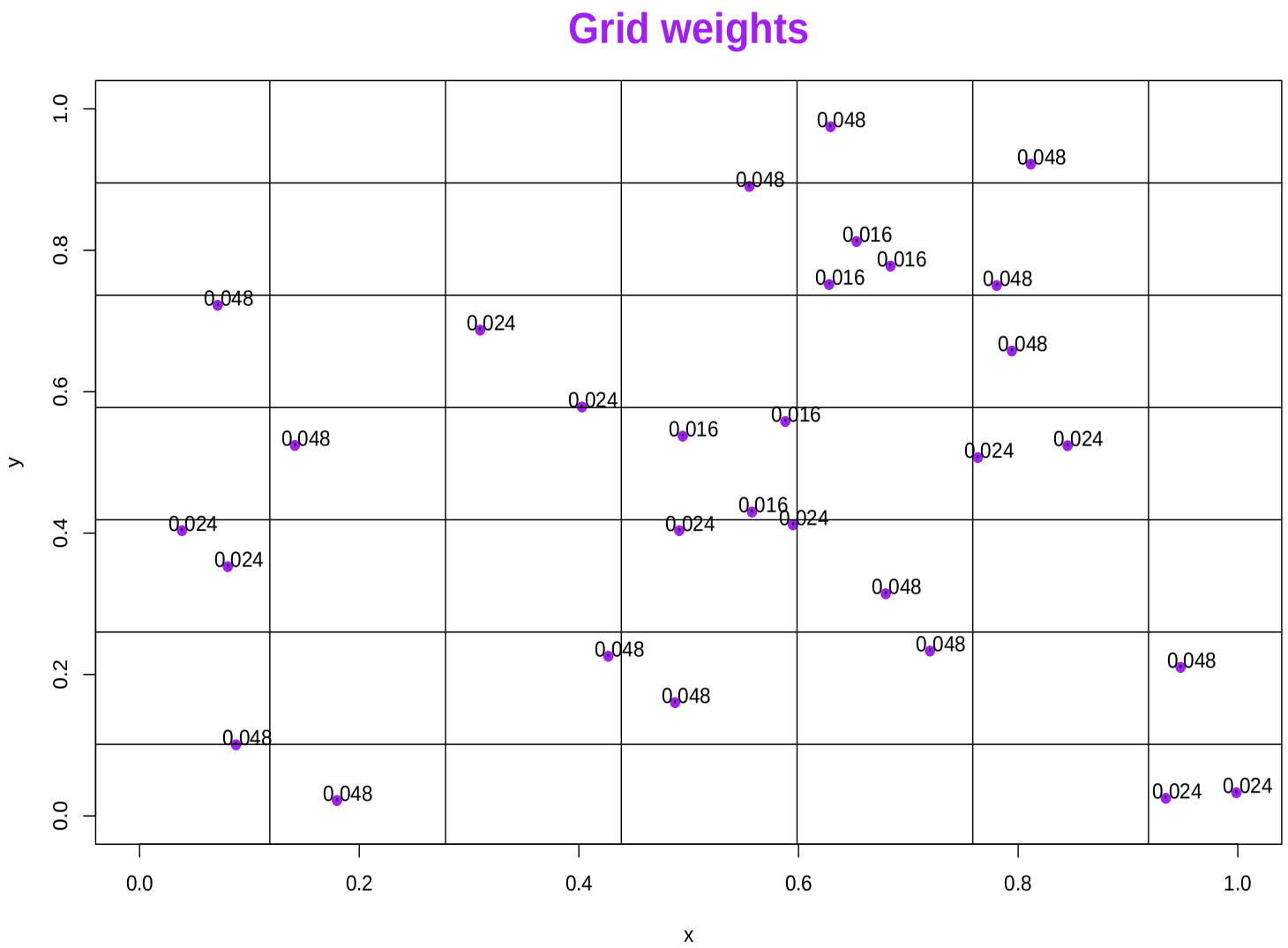
Possible weights:

- Voronoï tessellation
- constant for each square of a grid
- kriging weights

Voronoï tesselation



Grid method


Grid weights
0.048
0.048
0.048
0.016
0.016
0.016
0.048
0.048
0.024
0.048
0.024
0.016
0.016
0.048
0.024
0.024
0.024
0.016
0.024
0.024
0.048
0.048
0.048
0.048
0.048
0.048
0.024
0.024
y
x
0.0
0.2
0.4
0.6
0.8
1.0

Kriging Weights

Estimating the mean over a domain $\mathcal{D}$ of dependent data $(X_\alpha)_{\alpha=1,n}$
Best (minimal variance) Linear Unbiased Estimator

$$\widehat{\mu} = \sum_{\alpha=1}^{n} \lambda_\alpha X_\alpha \qquad \Lambda = \begin{pmatrix} (\lambda_\alpha)_{\alpha=1,n} \\ m \end{pmatrix}$$

with $\widetilde{C}\Lambda = \widetilde{c}$ and

$$\widetilde{C} = \begin{pmatrix} C & 1 \\ 1 & 0 \end{pmatrix} \qquad \widetilde{c} = \begin{pmatrix} c(x_\alpha, \mathcal{D}) \\ 1 \end{pmatrix}$$

$C_{\alpha,\beta} = \mathrm{Cov}(X_\alpha, X_\beta)$
$c(x_\alpha, \mathcal{D}) = \dfrac{1}{|\mathcal{D}|} \displaystyle\int_{\mathcal{D}} \mathrm{Cov}(x_\alpha, y)\mathrm{d}y.$

Kriging of the mean : $(\lambda_\alpha)_{\alpha=1,n}$ Kriging Weights
with the constraint $\lambda_\alpha > 0$

"natural declustering"

Spatial CART

p_{il} estimation of $P(Y = i \mid X \in B_l)$

$$P(Y = i \mid X \in B_l) = E(\mathbb{I}_{Y=i|X\in B_l})$$

estimated by kriging the mean of $\mathbb{I}_{Y=i|X\in B_l}$.

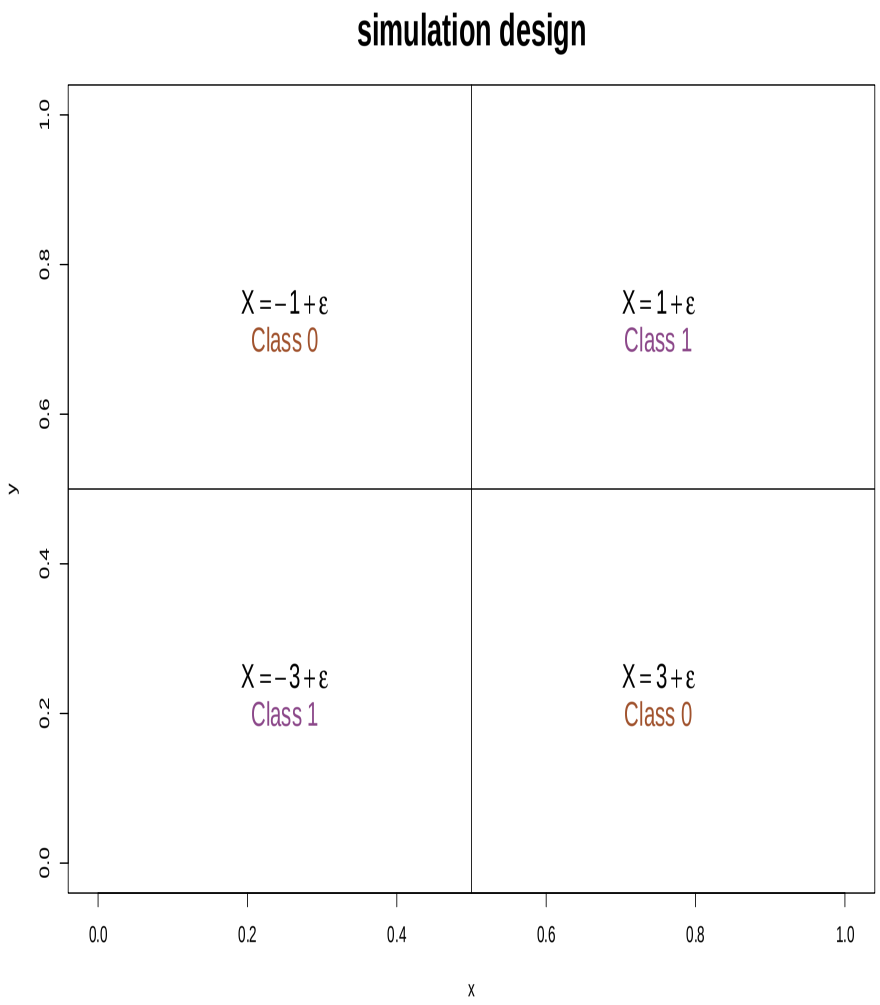
$$\widehat{p_{il}} = \sum_{\alpha: X_\alpha \in B_l} \lambda_\alpha \mathbb{I}(Y_\alpha = i)$$

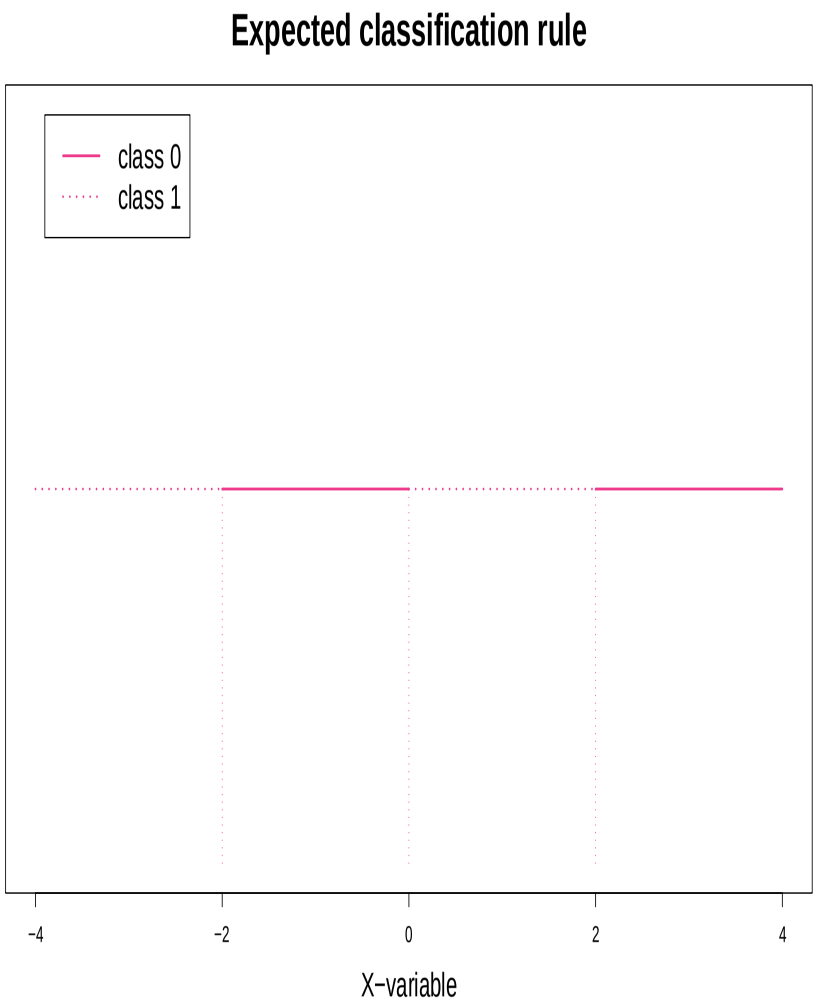
$R(T) = \mathsf{E}[\mathbb{I}_{T(X)\neq Y}]$ kriging the mean of $\mathbb{I}_{T(X)\neq Y}$.

$$\widehat{R}(T) = \sum_\alpha \lambda_\alpha \mathbb{I}(T(X_\alpha) \neq Y_\alpha)$$

Simulations

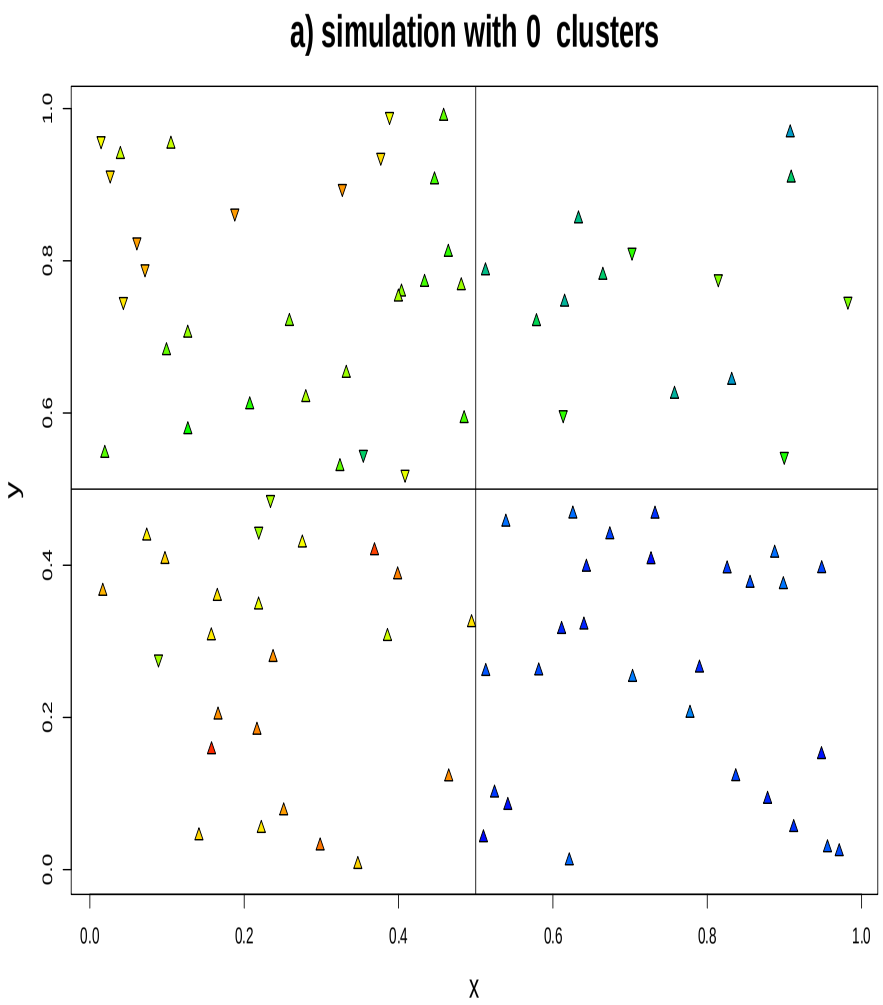
Simulation design

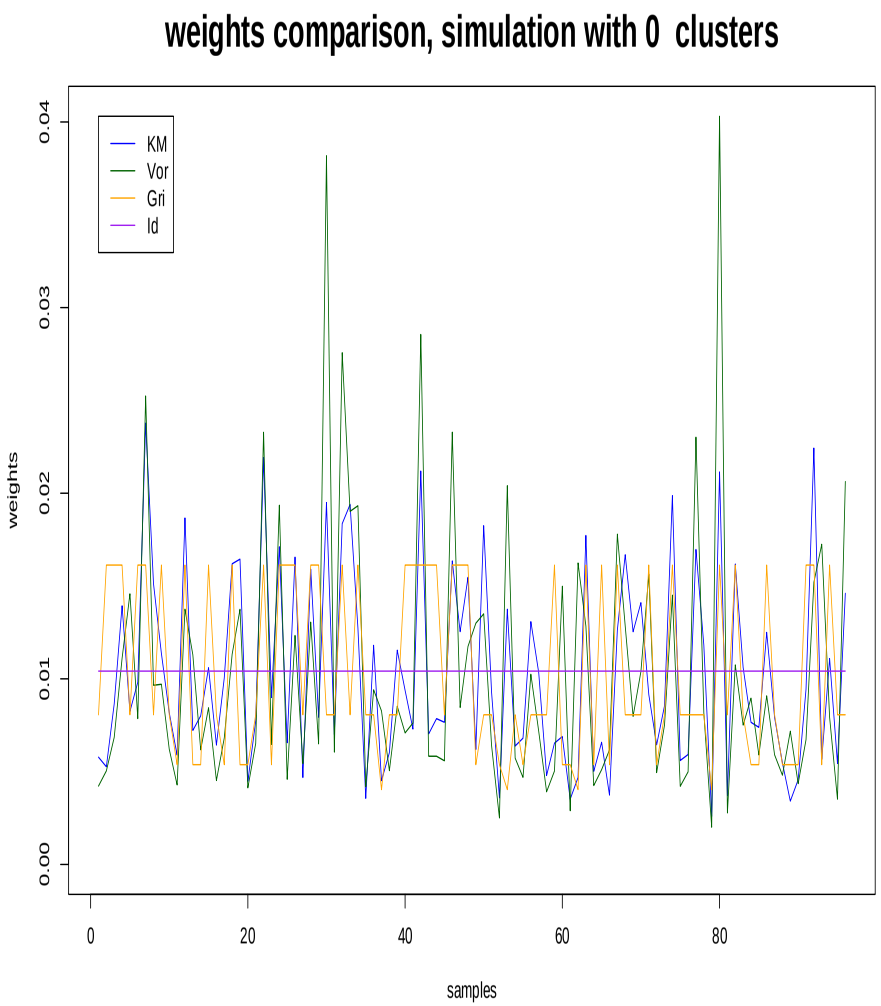




Simulations

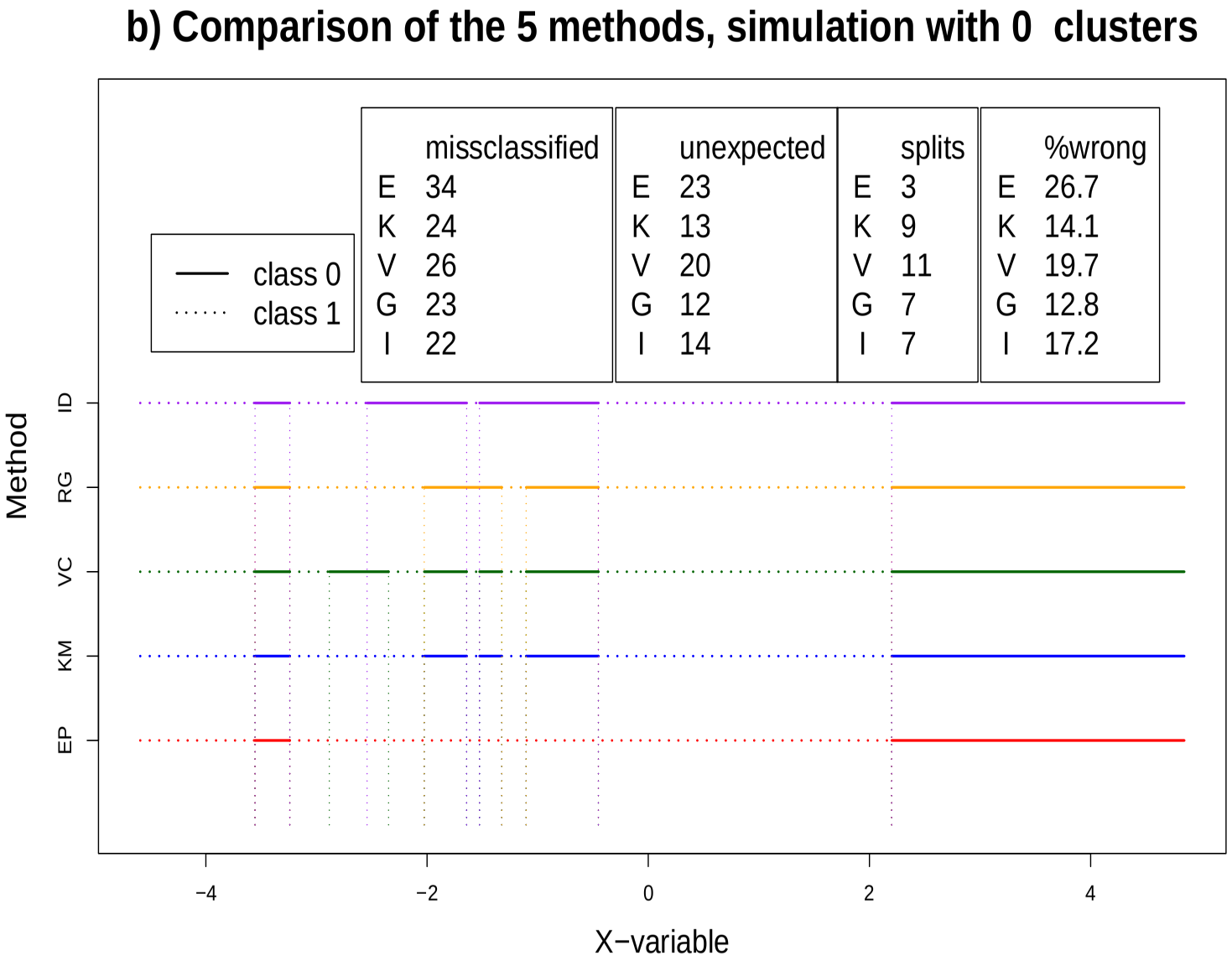
Simulation 0 clusters





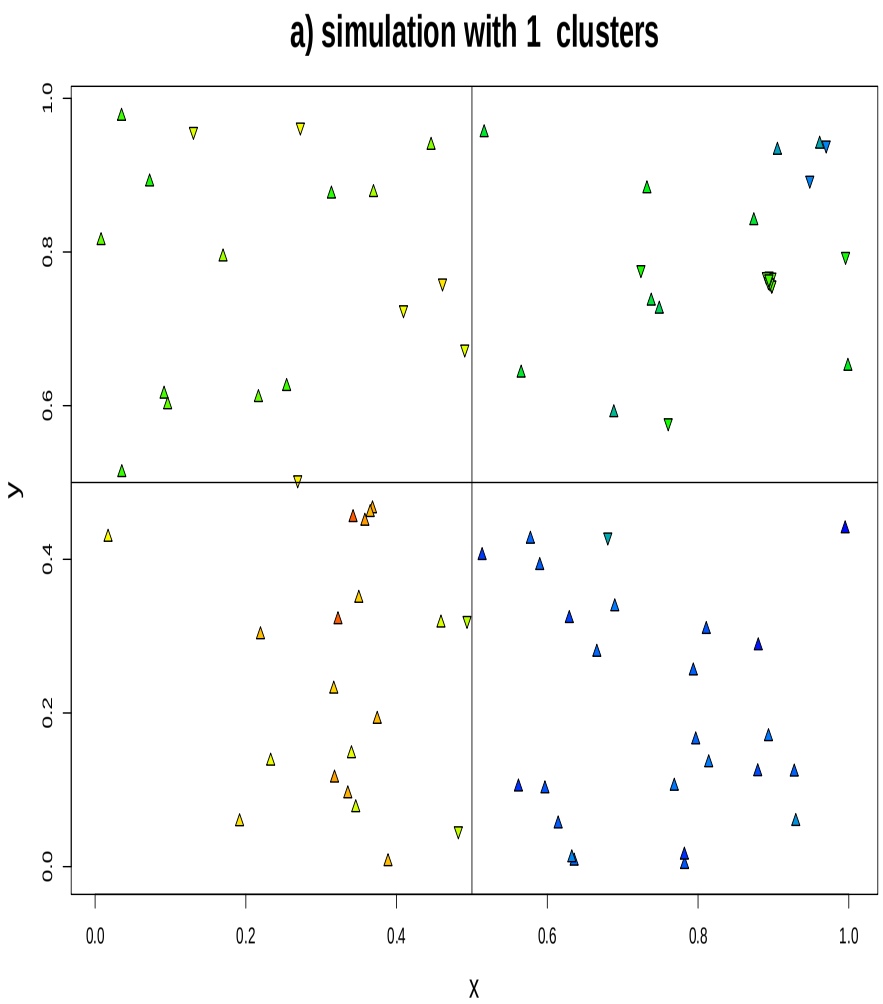
Simulations

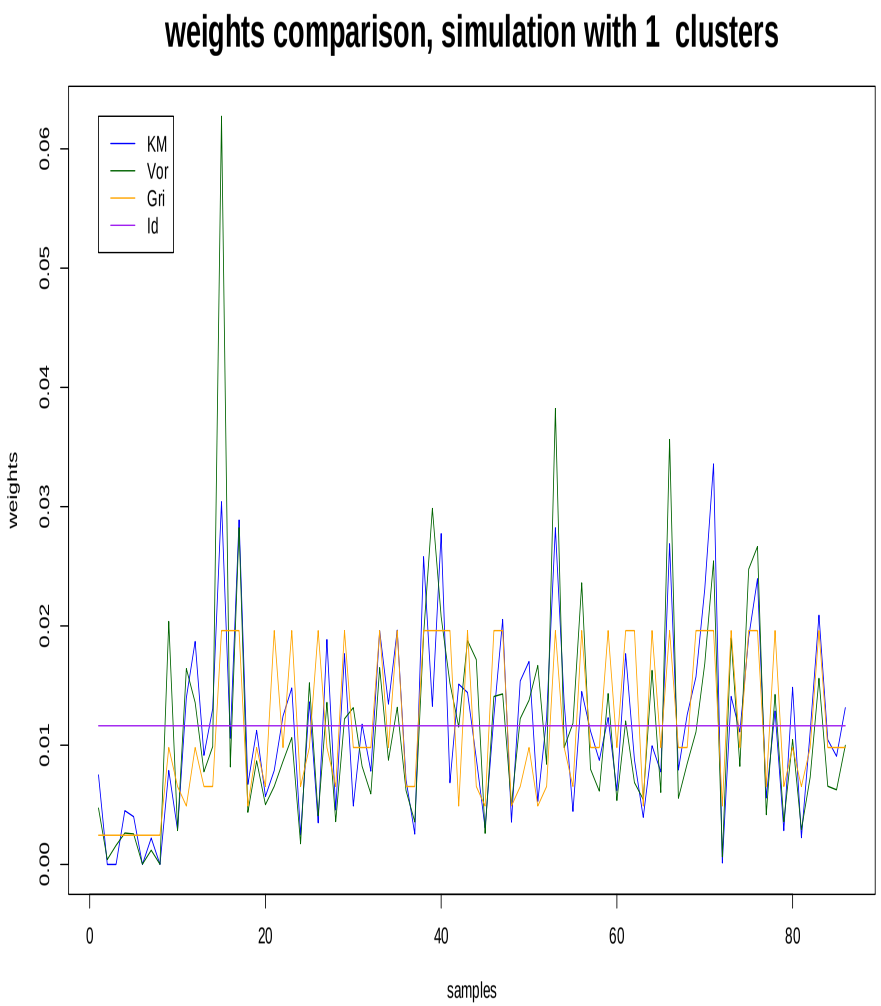
Simulation 0 clusters



Simulations

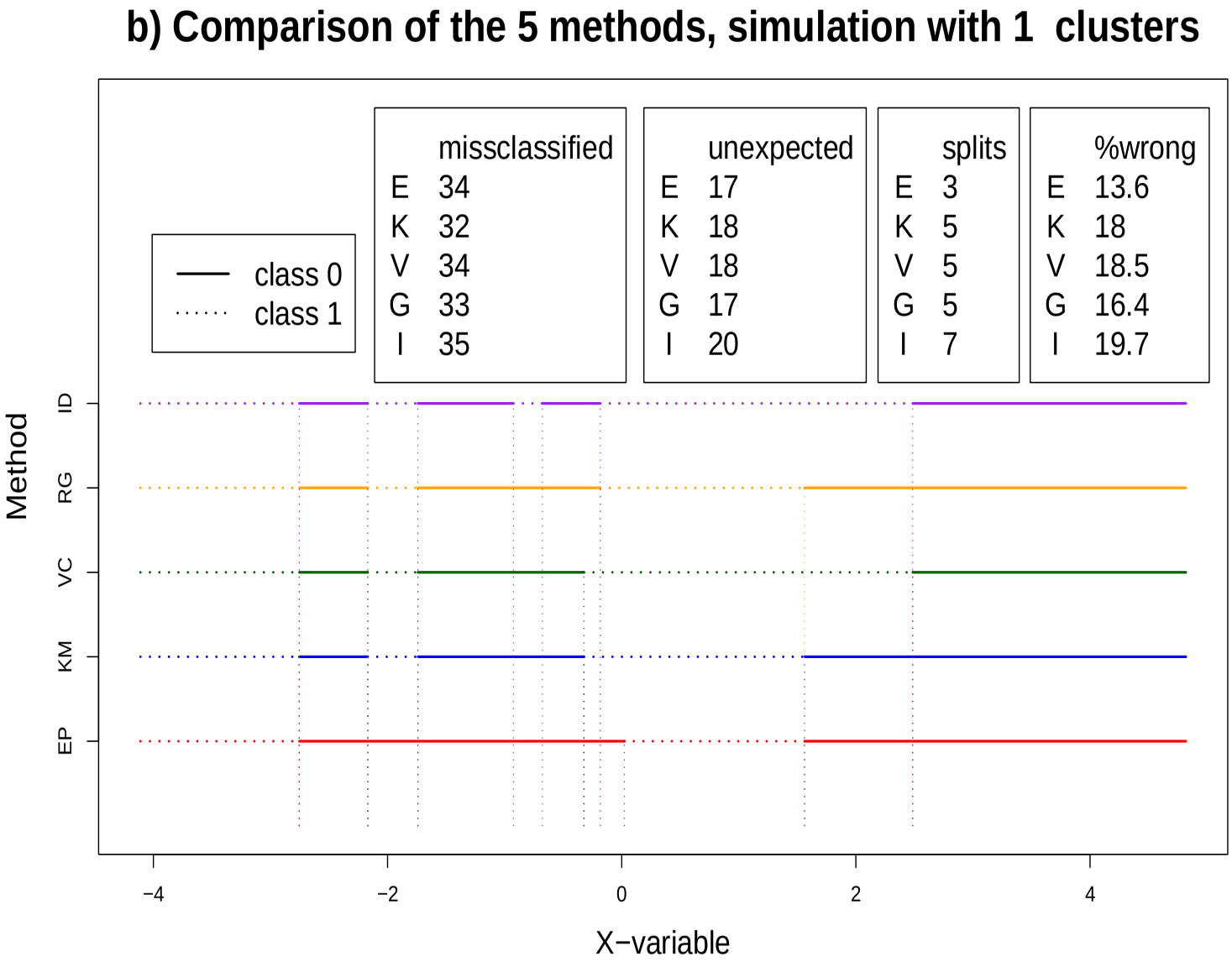
Simulation 1 clusters





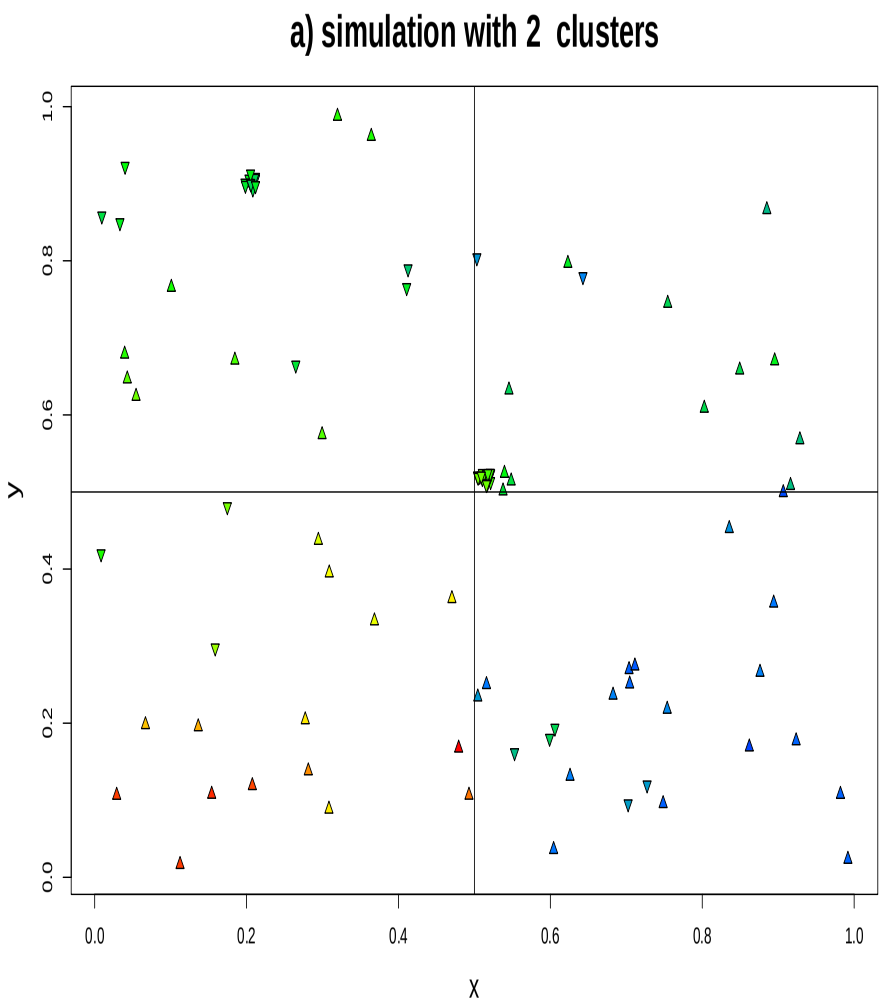
Simulations

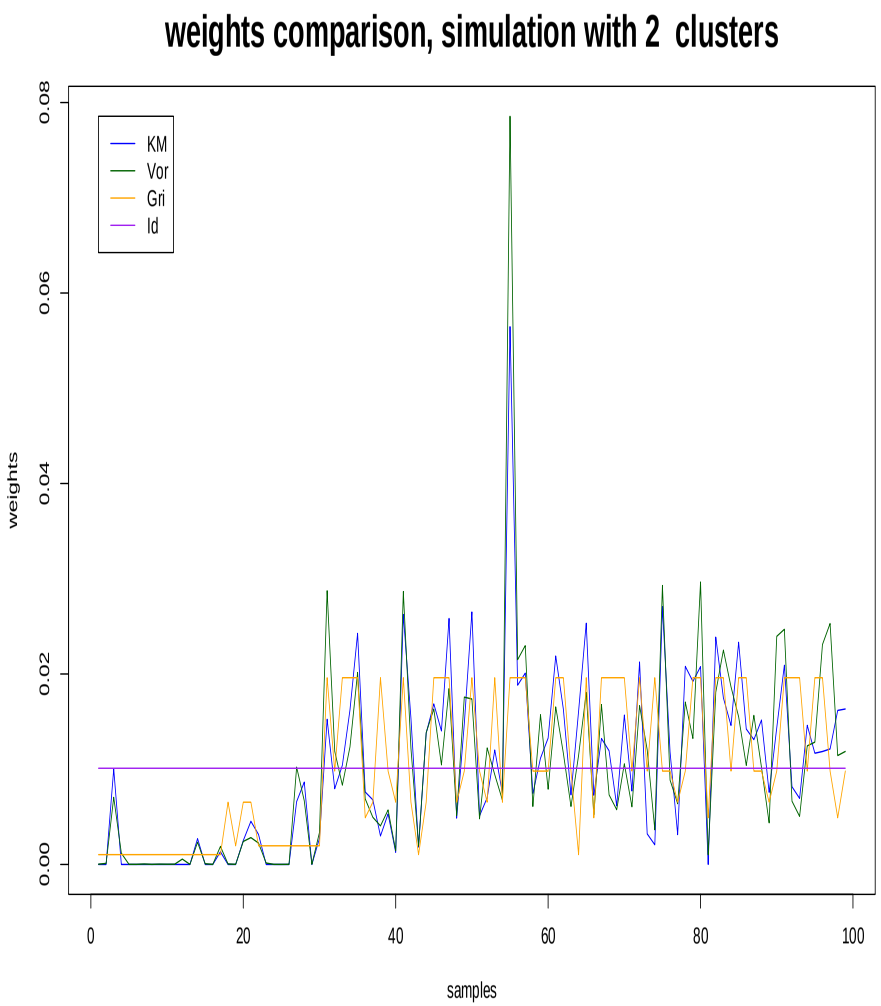
Simulation 1 clusters



Simulations

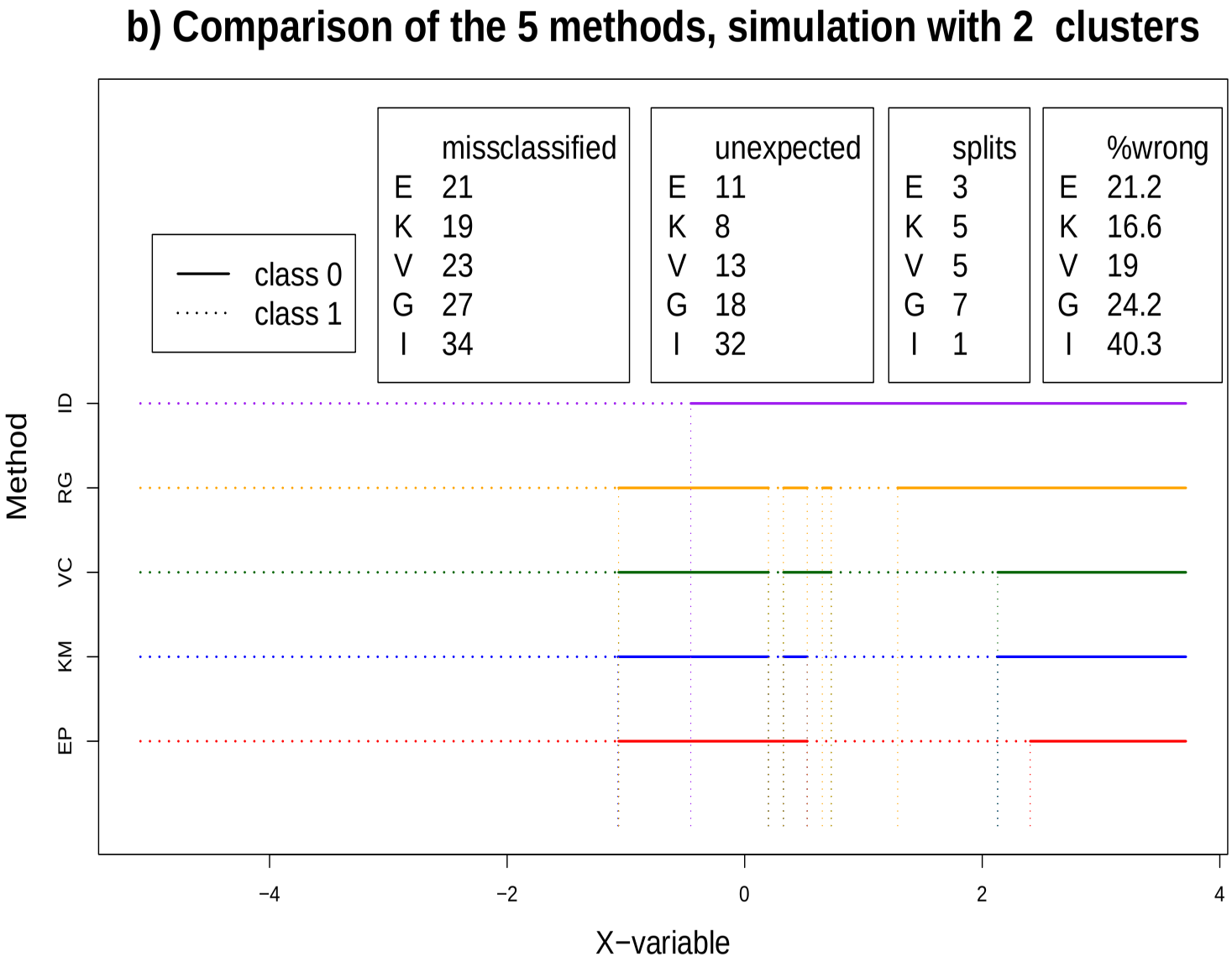
Simulation 2 clusters





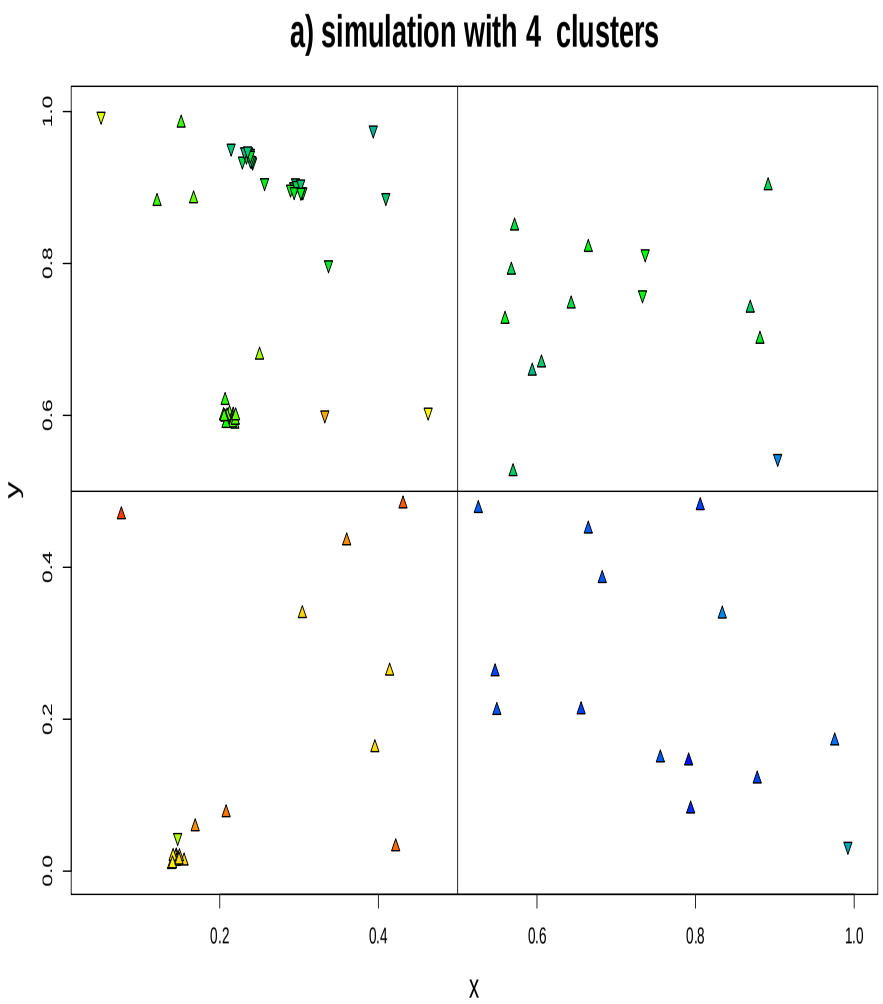
Simulations

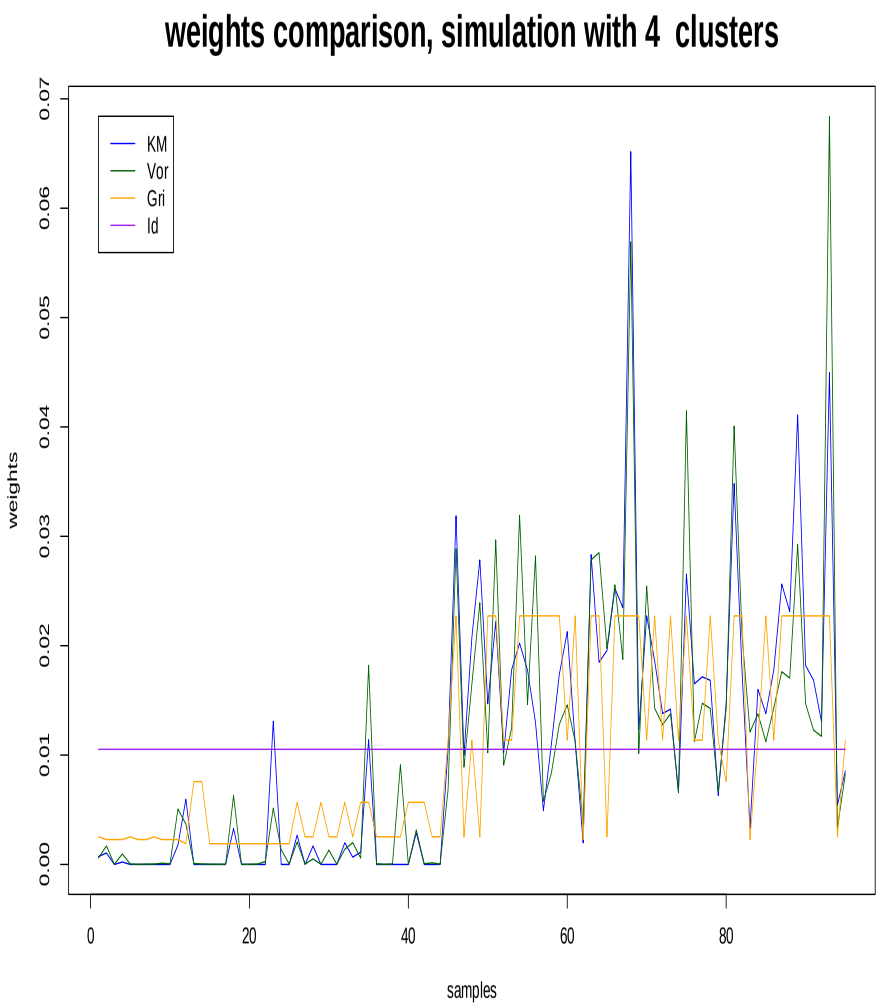
Simulation 2 clusters



Simulations

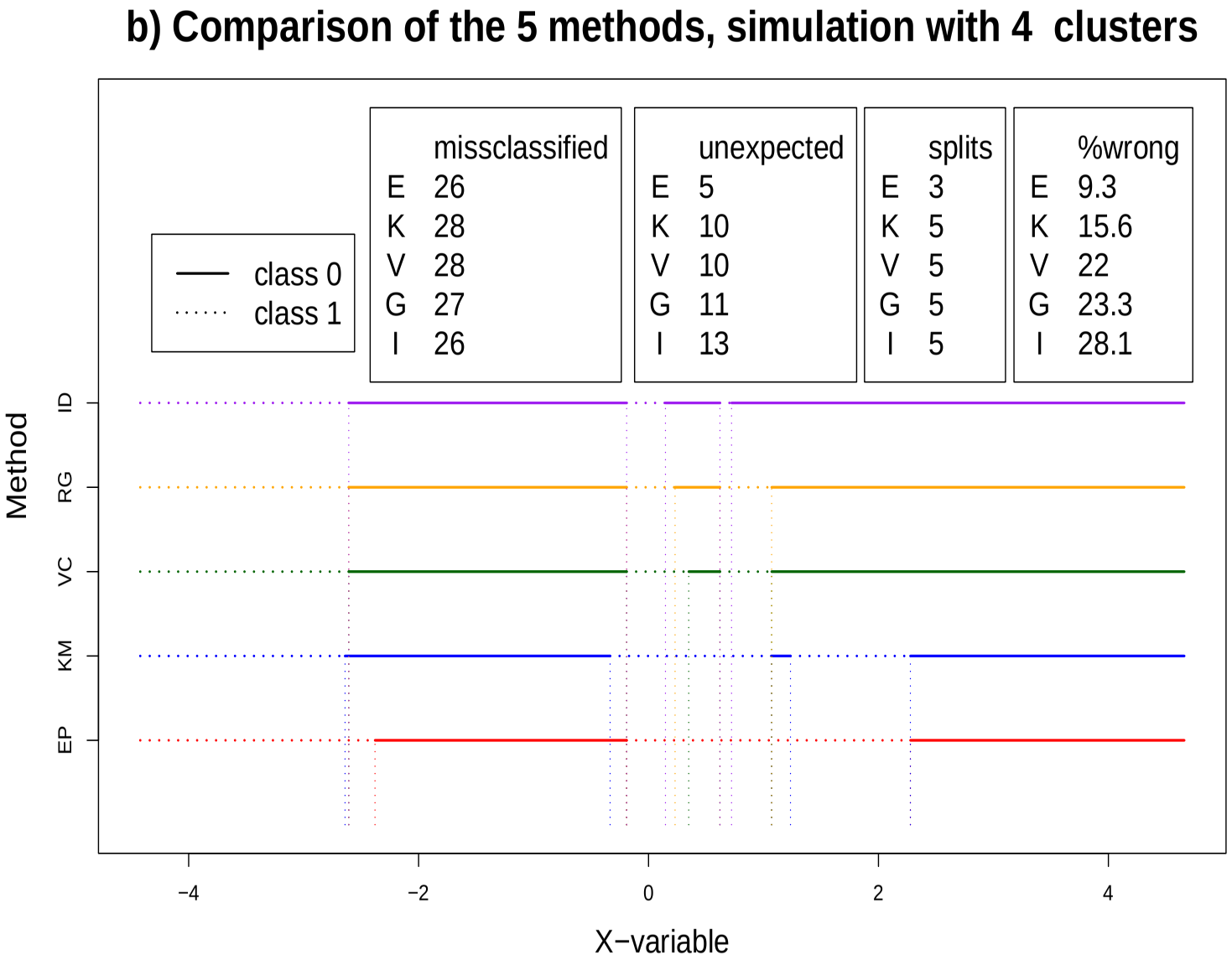
Simulation 4 clusters





Simulations

Simulation 4 clusters

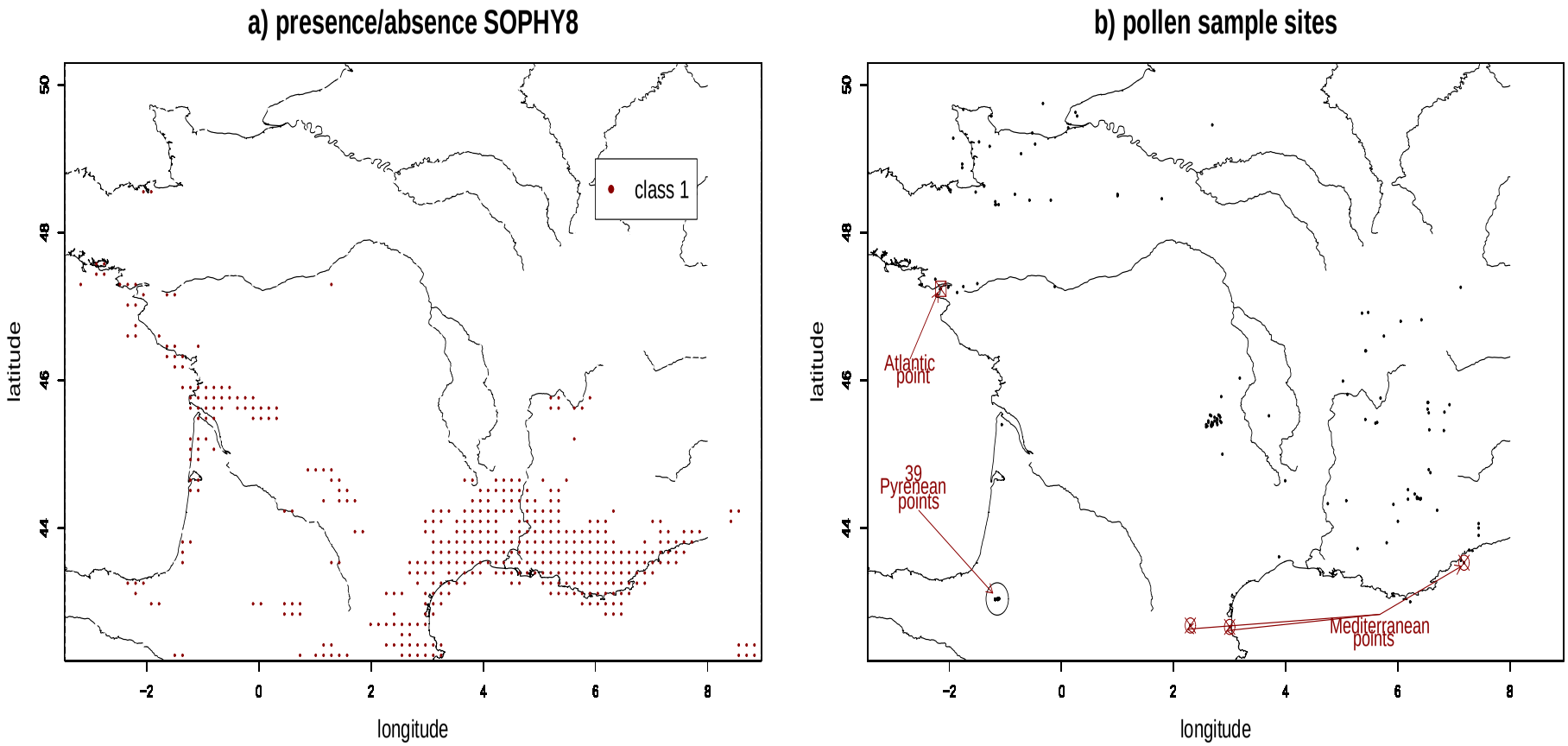


Simulations

Statistics on missclassified points on test samples.

clusters	EP		KM		Vor		Gri		Id	
	total	rule	total	rule	total	rule	total	rule	total	rule
0	**32.8**	**17.7**	29.9	12.7	30.4	13.3	30.3	11.9	**29.7**	**10.8**
1	28.3	14.5	27.7	12.1	**26.5**	**10**	27.8	12.8	**29.8**	**16.5**
2	28.3	14.1	25.9	**10.8**	26.9	11.7	**26.3**	**10.8**	**28.9**	**15.6**
≥ 3	29.3	14.4	**27.6**	11.3	27.9	**10.7**	28.5	11.3	**31.9**	**17.6**

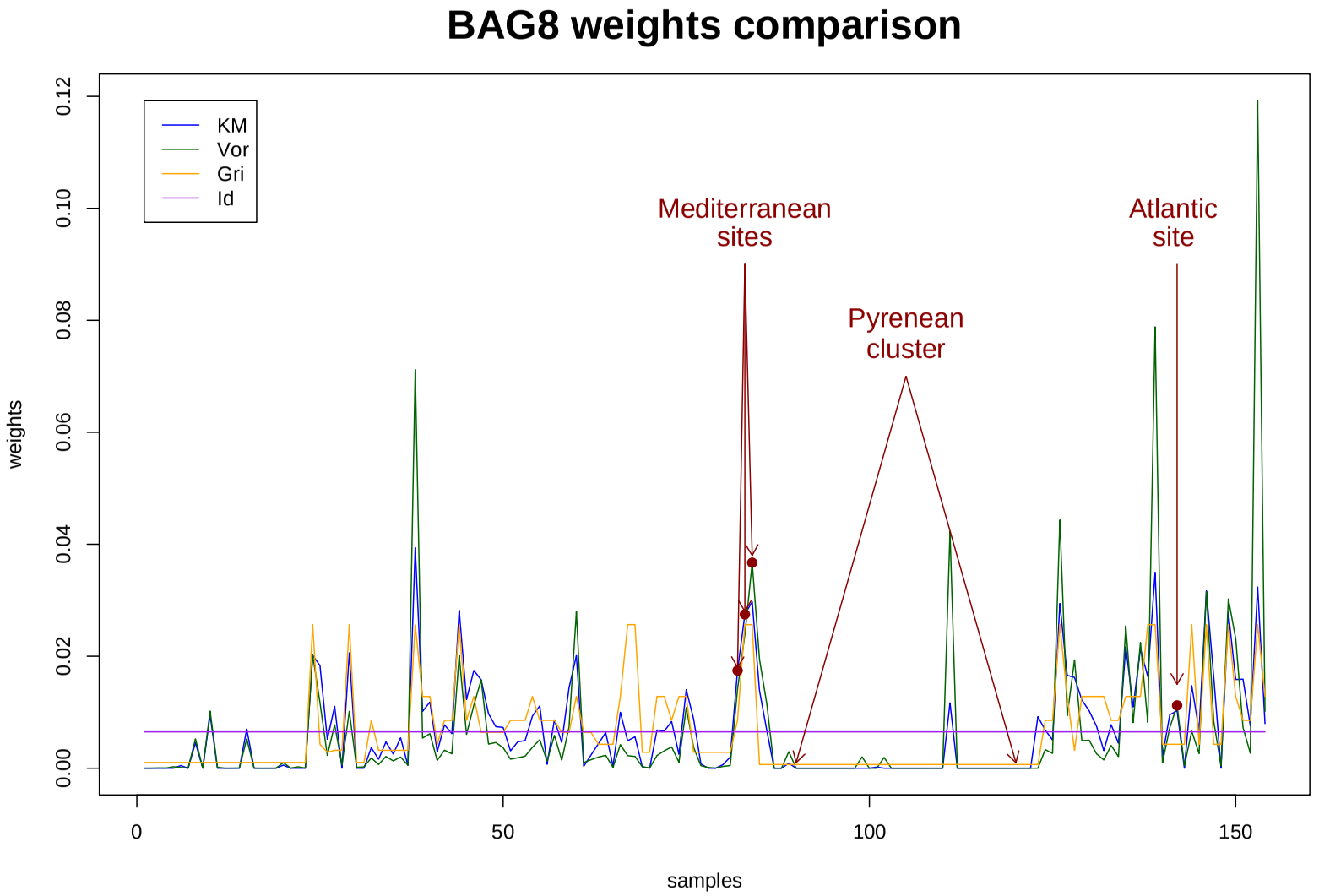
Ecological Data



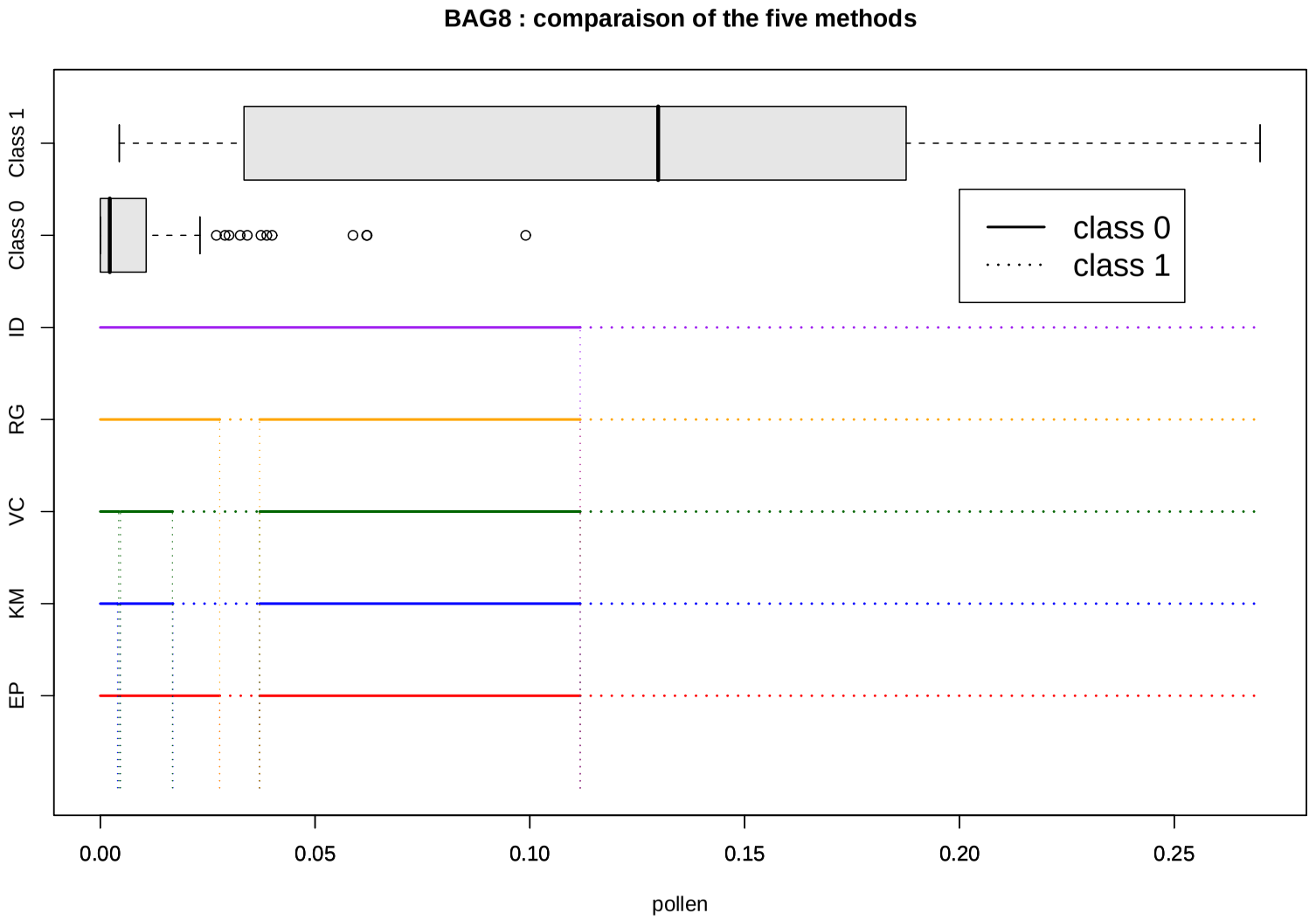
a) Presence and absence of specie from BAG8;

b) boxplots of pollen frequencies according to presence or absence

Weights


BAG8 weights comparison
KM
Vor
Gri
Id
Mediterranean sites
Pyrenean cluster
Atlantic site
weights
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
50
100
150
samples

Partitionning



Interpretation

- pollen frequency > 0.11, classified as presence for all methods.

- presence sites with low pollen frequency:
  - one at the border: large weight and well classified for Voronoï and Kriging the Mean.
  - 3 have similar pollen frequencies as the Pyrenean locations, but are isolated
    * Standard CART forget them
    * Spatial methods create a new cut

Conclusion

Weighted and Spatial CART

- discard clustered points of the analysis based on proximity or correlation properties
- provide weights easily interpretable
- highlight the effect of disturbance on the clustered points
- give a better understanding of the sampling scheme for the ecological data

Mostly useful when clustered data perturb the classification rule.