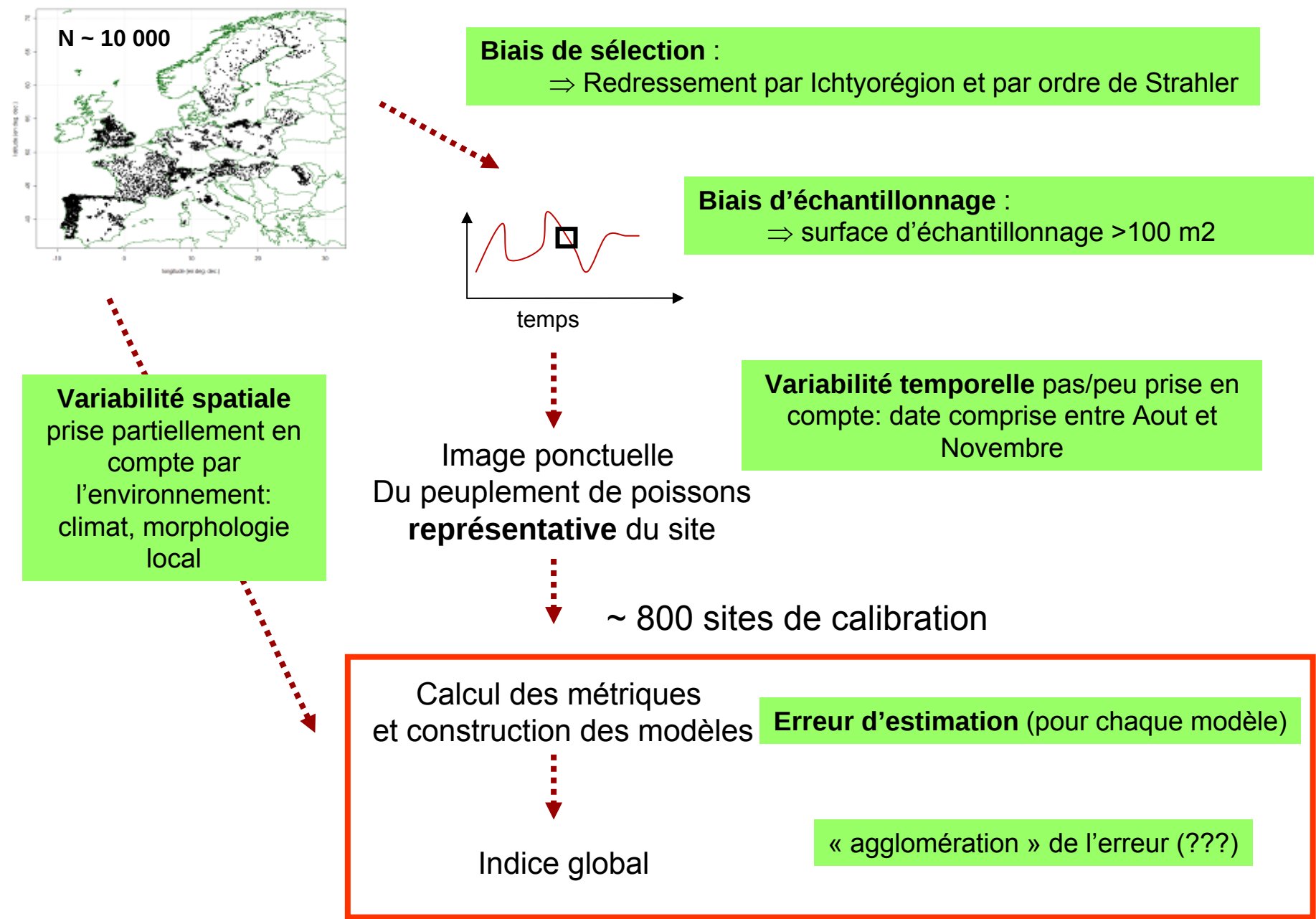

Incertitude & indice poisson :

Etat et perspectives



AQUAREF - PARIS 2008

Vue d'ensemble ultra-simplifiée



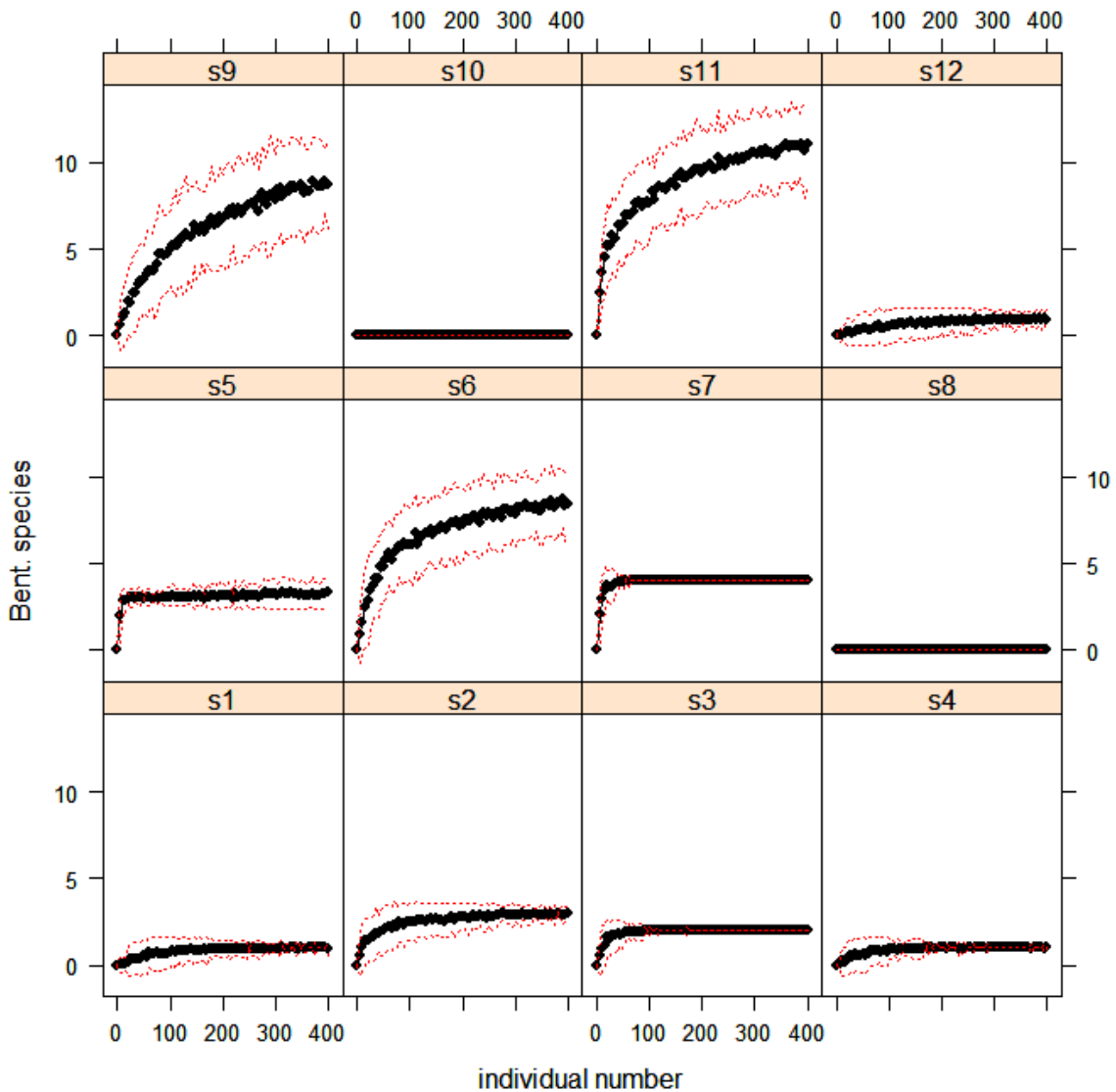
Incertitude et indices « Poissons »: petit aperçu du problème

Type d'incertitude		Réponses
Variabilité (variations naturelles)		
	Spatiale : plusieurs niveaux de variations (ex. local, global, structure du réseau)	- Prise en compte en partie dans les modèles (ex. taille du cours d'eau, climat) - Vérification <i>a posteriori</i> (ex. effets régionaux, auto-corrélation) - Redressement par Strahler et par Ichtyo-région
	Temporelle	- Pas prise en compte dans les modèles : le plus souvent les relevés sont ponctuels - On prend en compte seulement les opérations entre Aout-Novembre
Imprécision et erreur de mesure		
	Biais d'échantillonnage	- Limites de surface > 100m² - 50 captures minimum pour les sites de calibration - ratio longueur/largeur ? (cf DCE)
	Biais de sélection	- Redressement de l'échantillon par Strahler et par Ichtyo-région
	Erreur de mesures sur les variables environnementales	- Vérification des « grandes » relations (ex. distance à la source vs taille du BV) et exclusion des sites « douteux »
Estimation statistique		
	- Erreur d'estimation des modèles - Dispersion autour des réponses moyennes	- Hypothèses sur les lois de distribution des métriques - Intervalle de confiance et/ou de prédiction - Test de sensibilité (simulation) - Validation externe (jeu de données indépendants) et interne (bootstrap et CV)

Exemple: biais de mesure et erreur d'échantillonnage

Relation entre nombre de poisson et estimation du nombre d'espèces benthiques

Simulation basée sur une loi multinomiale



Test sur des sites de calibration
(quasi-référence)

$$\hat{S} = f(\text{individual number})$$

$$\hat{S} \pm \varepsilon_{1-0.05} \sqrt{\text{var}(\hat{S})}$$

$$\hat{S} = \sum_{i=1}^n \hat{S}_i / n$$

Développement d'une méthode commune à l'échelle de l'Europe (EFI+)

- ⇒ Généralisation de la Théorie de la Niche (Traits invariants, Convergence...)
- ⇒ Prise en compte de la variabilité naturelle par la Modélisation

1-Construction d'un modèle basé sur les sites calibrations

$$\text{Métrique} \sim \text{Environnement} + \varepsilon$$

Environnement:

- Climat
- Morphologie locale

2-Prédiction d'une valeur théorique basée sur le modèle

$$\text{Métrique théorique} \sim \text{Environnement}$$

3-Calcul de la distance entre observation et prédiction

$$\text{Métrique observé} - \text{Métrique théorique}$$

4-Agglomération des métriques

$$\text{Index} = f(\text{Métriques})$$

Pont, Hugueny, Beier, Goffaux, Melcher, Noble, Rogers, Roset & Schmutz (2006) Assessing river biotic condition at a continental scale: a European approach using functional metrics and fish assemblages. *Journal of Applied Ecology*, **43**(1), 70-80.

Pont, Hugueny & Rogers (2007) Development of a fish-based index for the assessment of river health in Europe: the European Fish Index. *Fisheries Management and Ecology*, **14**(6), 427-39.

Le modèle linéaire (LM) et intervalles de confiance et de prédiction

Régression linéaire multiple (lien identité, distribution gaussienne):

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m + \varepsilon$$

Intervalle de confiance ou de prédiction :

$$\hat{y}_x \pm \hat{\sigma}(\hat{y}_x) \cdot t_{1-\alpha, n-p} \quad \text{avec} \quad \hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

1) Estimation de la variance pour les intervalles de **confiances** :

$$\hat{\sigma}^2(\hat{y}_x) = \hat{\sigma}^2 \mathbf{X}_x^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}_x$$

2) Estimation de la variance pour les intervalles de **prédictions** :

$$\hat{\sigma}^2(\hat{y}_x) = \hat{\sigma}^2 \left(1 + \mathbf{X}_x^t (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}_x \right)$$

Le modèle linéaire généralisé (GLM) et intervalles de confiance et de prédiction

Les **GLM** sont définis par les trois composantes suivantes :

1) Fonction de distribution de la famille **exponentielle**

2) Prédicteurs linéaires :

$$g(y) = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m + \varepsilon$$

3) Fonction de **lien** $g()$:

$$E(Y) = \mu = g^{-1}(\eta)$$

1) Erreur dans l'espace du lien :

Calcul de l'intervalle de confiance (ou de prédiction) au niveau de la fonction additive (voir LM)

2) Méthode delta (résultats asymptotique) :

Intervalle de confiance (ou de prédiction) symétrique

$$\text{var}(\hat{\mu}) = \left(\frac{\delta\mu}{\delta\eta} \right)^2 \text{var}(\hat{\eta}) \quad \text{avec} \quad \text{var}(\hat{\eta}) = \mathbf{X} \text{var}(\hat{\beta}) \mathbf{X}^t$$

3) Report de l'erreur :

3.1) Calcul de l'intervalle de confiance (ou de prédiction) au niveau de la fonction additive.

3.2) Utilisation de la transformation (inverse du lien du modèle).

$$IC(\mu) = g^{-1}(IC(\eta))$$

Procédure de calcul utilisé dans les indices poissons (FAME et EFI+)

1) Pour une métrique donnée:

$$y_{obs} \quad , \quad y_{pred} \quad \text{et} \quad \left[y_{\min} = y_{pred} - \Delta_{-}y ; y_{\max} = y_{pred} + \Delta_{+}y \right]$$

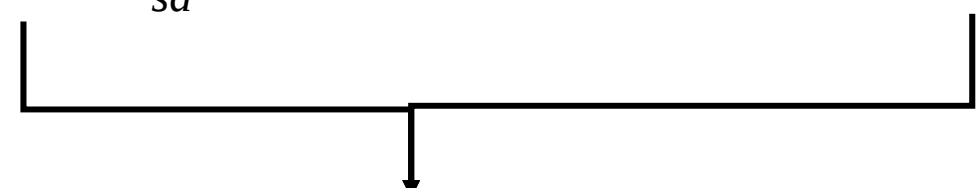
2) Distance entre observation et prédiction pour chaque métrique

$$dist(obs, pred) = \frac{(y_{obs} - y_{pred}) - \bar{m}}{sd}$$

← moyenne des scores de calibration

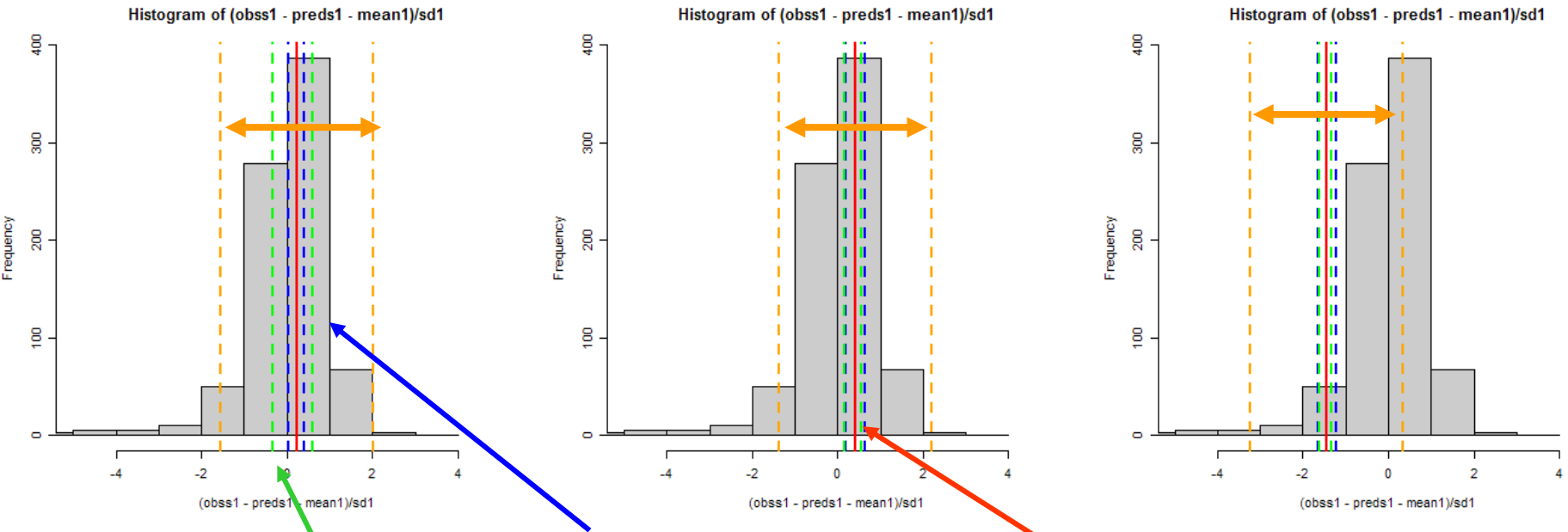
← Std dev. des scores de calibration

3) Distance entre observation et limites de l'intervalle

$$dist(obs, \min) = \frac{(y_{obs} - y_{\min}) - \bar{m}}{sd} \qquad \qquad \qquad dist(obs, \max) = \frac{(y_{obs} - y_{\max}) - \bar{m}}{sd}$$

$$[dist(obs, \min); dist(obs, \max)]$$

Exemple pour le score basé sur le nombre d'espèces intolérantes à des faibles concentrations en O2 ("ric.WQO2.O2INTOL"):

↔ Intervalle de prédiction à 90%



Intervalle de confiance à 90%

Intervalle estimé par bootstrap basé sur les premiers et derniers percentiles (Davidson & Hinkley 1997)

Distance standardisée entre observation et prédiction

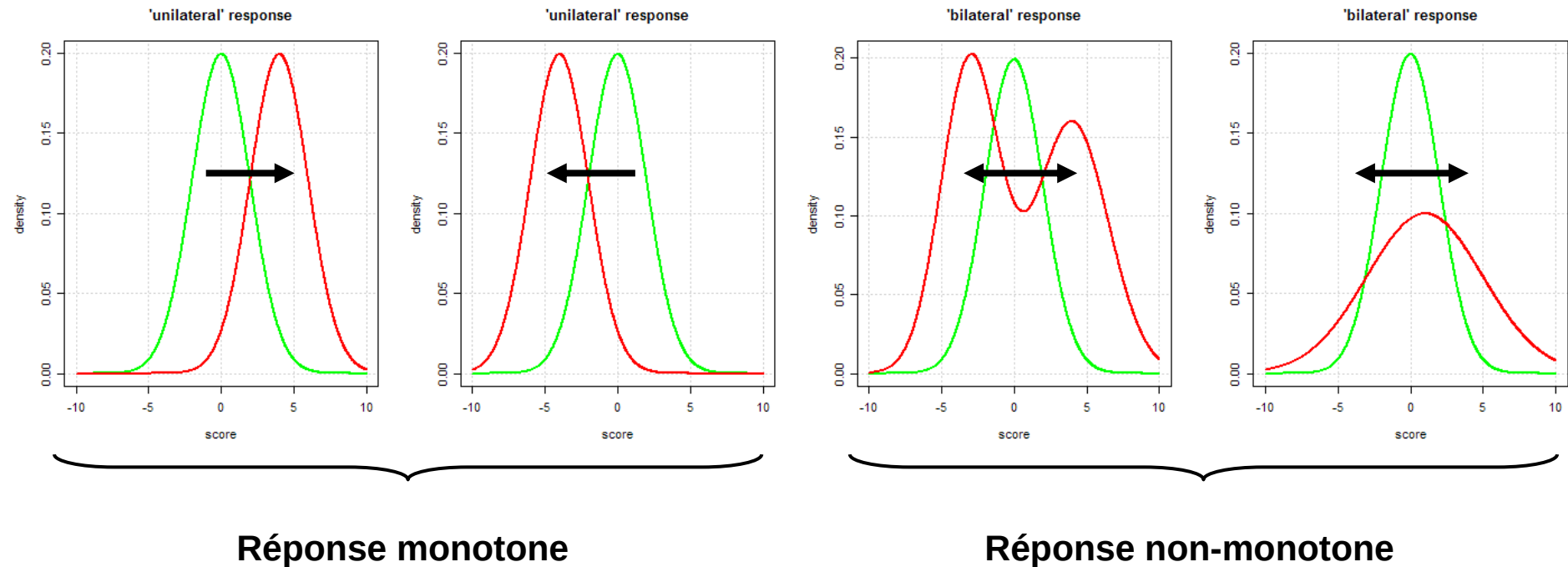
Transformation de la métrique en un score contenu dans [0,1]

1) Report des valeurs extrêmes

- ⇒ Nécessite l'utilisation d'une fonction monotone (**vrai** pour les réponses unilatérales, **faux** pour les réponses bilatérales)
- ⇒ Cherché un facteur correcteur pour la continuité dans le cas de réponse bilatérale

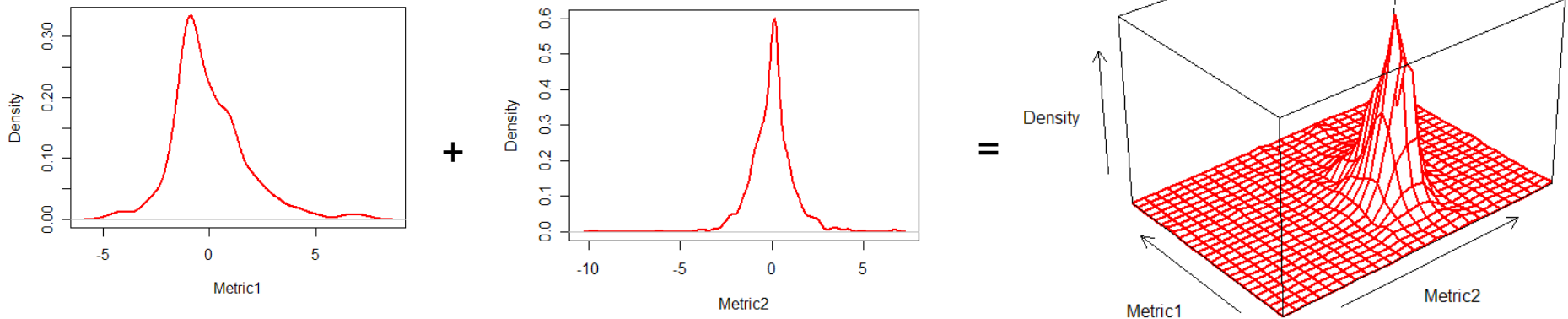
2) Méthode de type bootstrap (tendance à l'optimisme)

Type de réponse :



Agglomération des métriques

Exemple d'association entre 2 métriques :



Loi normale multi-dimensionnelle de variables **non-indépendante** (!?)

1) par défaut, moyenne des scores (approche FAME)

2) distance de Mahalanobis:
$$d_i^2 = (x_i - \mu)^t \Sigma^{-1} (x_i - \mu)$$

Le calcul d'un intervalle de tolérance est compliqué dans tous les cas !



En cours d'étude ...

Conclusion ...

- 1) **Validation** externe et interne des modèles pour limiter les biais d'estimation des valeurs prédites (ex. bootstrap, cross-validation)
- 2) Prévoir des analyses de sensibilité pour tester les modèles (**crash-test**)
- 3) L'erreur de mesure/estimation et variabilité naturelle sont deux processus différents.
- 4) Pour la construction des « fameuses » 5 classes, on a besoin de définir les critères de découpage (**fonction de coût ***).
- 5) Informer l'utilisateur des **limites** et des risques de **mauvais usages** de l'outil (ex. « range » d'utilisation).