Analyse de sensibilité et métamodélisation basées sur les polynomes de chaos et la régression PLS pour des modèles numériques à entrées corrélées

Rencontres Mexico-MascotNum, IFREMER, Nantes 23-24 Nov, 2016

Jean-Pierre Gauchi (MaIAGE / Jouy)
en collaboration avec : A. Bensadoun (postdoc MaIAGE) , F. Colas (doctorante EA), N. Colbach (INRA / EA / EcolDur / Dijon)

Plan de l'exposé

- Rappel contexte entrées indépendantes
- Rappel contexte polynômes de chaos
- Influence de la corrélation
- Nouvelle proposition pour entrées continues corrélées
- Rappel régression PLS
- Nouveaux indices de sensibilité
- Application à deux modèles académiques
- Application au modèle numérique FlorSys
- Améliorations
- Perspectives

Rappel contexte entrées indépendantes - 1

- On considère le modèle mathématique:

$$\begin{aligned} f : \mathbb{R}^p &\longrightarrow \mathbb{R} \\ \mathbf{X} &\longmapsto Y = f(\mathbf{X}) \end{aligned}$$

où:

  - $\mathbf{X} = (X_1, ..., X_p)$, vecteur à p composantes X_j, $j = 1..., p$ (variables explicatives = entrées=inputs),
  - f est une fonction déterministe, de forme explicite ou non, qui peut être très compliquée (systèmes d'équations différentielles, boîte noire).

- On suppose que pour tout point $\mathbf{x}_i = (X_{i1}, ..., X_{ip}) \in \Xi \subset \mathbb{R}^p$, on obtient toujours par un code informatique (le modèle numérique M) un résultat y_i (output) pour ce modèle:

$$y_i = f(\mathbf{x}_i) = M(\mathbf{x}_i)$$

Rappel contexte entrées indépendantes - 2

- Soit Ξ l'hypercube unité, la décomposition fonctionnelle ANOVA de Hoeffding (HDMR) consiste à exprimer f comme une somme de 2^p fonctions de dimension croissante (où f_0 est une constante):

$$f(\mathbf{X}) = f_0 + \sum_{j=1}^{p} f_j(X_j) + \sum_{1 \leq j < k \leq p} f_{jk}(X_j, X_k) + \ldots + f_{12\ldots p}(X_1, \ldots, X_p)$$

- Cette décomposition est **unique** si et seulement si:

$$\int_0^1 f_{j_1,\ldots,j_s}(X_{j_1,\ldots,j_s}) dX_{j_k} = 0 \quad \forall k = 1, \ldots, s, \quad \forall \{j_1, \ldots, j_s\} \subseteq \{1, \ldots, p\}$$

- Une conséquence: les fonctions $f_{j_1,\ldots,j_s}$ sont deux à deux **orthogonales** (si au moins un des indices n'est pas répété):

$$\int_0^1 f_{j_1,\ldots,j_s}(x_{j_1,\ldots,j_s}) f_{k_1,\ldots,k_t}(x_{k_1,\ldots,k_t}) dx = 0$$

$\Longrightarrow$ DONC: toute fonction de carré intégrable admet une décomposition fonctionnelle ANOVA unique si ses variables sont définies sur l'hypercube unité avec la propriété importante que les fonctions sont deux à deux $\perp$

Rappel contexte entrées indépendantes - 3

Supposons maintenant que $\mathbf{X}$ est un vecteur aléatoire $\sim \mathcal{U}\left([0,1]^p\right)$, alors:

1. On obtient la décomposition de la variance de la sortie aléatoire Y:

$$V(Y) = \sum_{j=1}^{p} V_j + \sum_{1 \leq j < k \leq p}^{p} V_{jk} + \ldots + V_{1\ldots p}$$

2. On définit les indices de sensibilité de Sobol:

$$S_j = \frac{V\left(E\left[Y|X_j\right]\right)}{V(Y)} \quad \text{et} \quad S_{T_j} = \frac{E_{\mathbf{X}_{-j}}\left[V_{X_j}\left(Y|\mathbf{X}_{-j}\right)\right]}{V\left(Y\right)}$$

Rappel contexte entrées indépendantes - 4

Estimation des indices, parmi de nombreuses méthodes on peut citer

- Méthodes de **Monte Carlo**: Sobol, Saltelli, Tarantola, ...===> LHS, Quasi-Monte-Carlo (points $\mathbf{x}_i$ à discrépance faible), par exemple la suite LP_τ de Sobol.
- Méthodes **FAST** (et EXTENDED FAST): décomposition spectrale de Fourier de $f(\mathbf{X}) \Longrightarrow$ réduction du coût de calcul des intégrales multidimensionnelles en les remplaçant par une intégrale unidimensionnelle.
- Par **métamodélisation** préalable: processus gaussiens, RKHS (décomposition de type ANOVA pour un choix approprié de noyaux), splines, etc.
- **Polynomes de chaos**: toute variable de carré intégrable peut s'écrire sous la forme d'un développement en polynomes orthogonaux (pour une mesure particulière) $\Longrightarrow$ estimation des indices par régression.

Rappel contexte polynomes de chaos - 1

- Soit Z une v.a. de second ordre, on défnit (Wiener,1938; Cameron, 1947) la série du chaos polynômial:

$$Z = \sum_{k=0}^{\infty} Z_k \Psi_k \left(\{\xi_p\}_{p=1}^{\infty} \right)$$

où:

  - les Z_k sont des coefficients réels,
  - $\left\{\xi_p\right\}_{p=1}^{\infty}$ une suite infinie de v.a. $\mathcal{N}(0,1)$ indépendantes,
  - les $\Psi_k(.)$ sont des polynômes orthogonaux (par rapport à la mesure de Gauss) multivariés de Hermite.

- Cette relation se généralise pour une mesure uniforme avec des polynômes orthogonaux multivariés de Legendre.

Rappel contexte polynomes de chaos - 2

- En pratique, on tronque la série infinie pour construire une approximation $M_d(\mathbf{X})$ de la sortie Y d'un code numérique, dépendant de p entrées, en choisissant un degré d :

$$Y \approx M_d(\mathbf{X}) = \sum_{k=0}^{P} \beta_k \Psi_k(\mathbf{X}) \ , \qquad \mathbf{X} \sim \mathcal{U}\left([-1,+1]^p\right)$$

- Chaque fonction $\Psi_k(\mathbf{X})$ d'un monôme, centrée par construction, élément d'un polynôme orthogonal multivarié de Legendre, est défini par une fonction spécifique des termes $X_1^{d_1}, X_2^{d_2}, ..., X_p^{d_p}$, avec $0 \leq d_j \leq d$, $\sum_{j=1}^{p} d_j \leq d$ ($P = [(p+d)!/(p!d!)] - 1$).
- MCO pour calculer $\beta \Longrightarrow \hat{\beta}^{[OLS]} \Longrightarrow$ bonne approximation des indices de Sobol à partir d'un certain d (Sudret, 2008).

Rappel contexte polynomes de chaos - 3

Grâce à la décomposition orthogonale de la variance on obtient P indices de sensibilité individuels, $IMSI_k$, $k = 1, ..., P$, définis par:

$$IMSI_k = \left(\hat{\beta}_k^{[OLS]}\right)^2 E\left\{\Psi_k^2(\mathbf{X})\right\} / D_{PC}$$

où D_{PC} is définie par:

$$D_{PC} = \sum_{k=1}^{P} \left(\hat{\beta}_k^{[OLS]}\right)^2 E\left\{\Psi_k^2(\mathbf{X})\right\}$$

Rappel contexte polynomes de chaos - 4

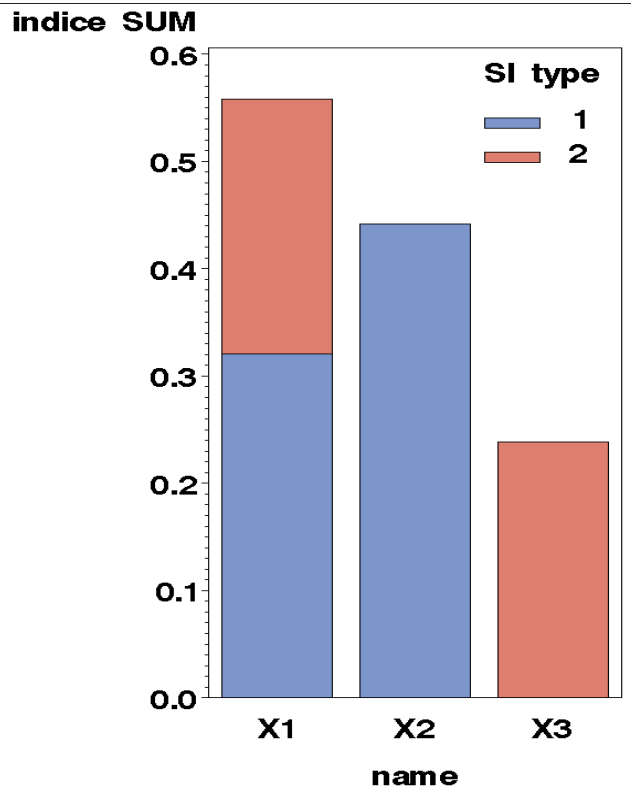
Trois types d'indices de sensibilité peuvent être définis pour chaque entrée X_j:

- Un indice qui représente l'**effet linéaire** : *LSI* (linear sensitivity index): $IMSI_k$ pour $k = j$.
- Un indice qui représente l'**effet polynômial** : *PSI* (polynomial sensitivity index) obtenu par sommation des $IMSI_k$ calculés à partir de X_j, $X_j^2, ..., X_j^d$.
- Un indice qui représente l'**effet total** : *TSI* (total sensitivity index), obtenu par sommation de tous les $IMSI_k$ calculés à partir de termes où X_j est présent: par exemple pour $d = 3$ et $p = 3$ on a $P = 19$, et 7 $IMSI_k$ où X_1 est présent, donc TSI_1 est la somme des $IMSI_k$ correspondant aux termes $\Psi_1(X_1)$, $\Psi_2(X_1^2)$, $\Psi_3(X_1^3)$, $\Psi_4(X_1X_2)$, $\Psi_5(X_1X_3)$, $\Psi_6(X_1X_2^2)$, $\Psi_7(X_1X_3^2)$.

Rappel contexte polynomes de chaos - 5

Ex Ishigami (entrées U)

$$Y = \sin(X_1) + 7[\sin(X_2)]^2 + 0.1X_3^4 \sin(X_1)$$



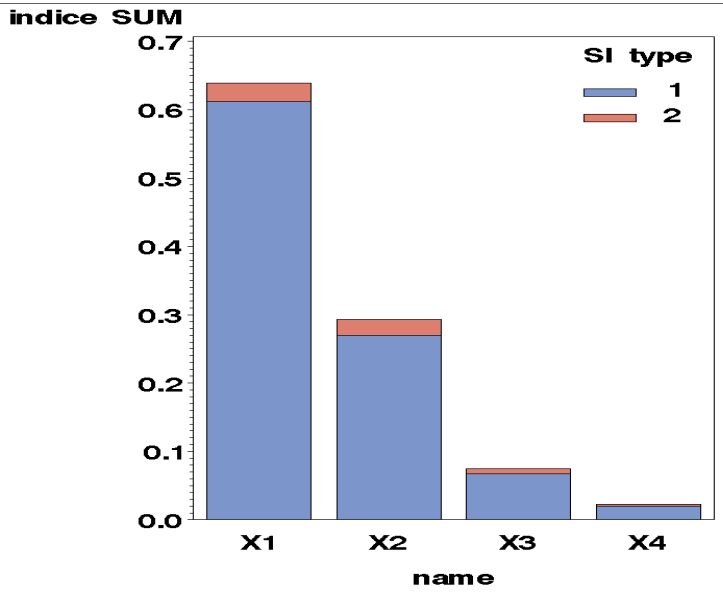
Les indices théoriques :

$X1$: 0.3138 / 0.5574 $X2$: 0.4424 / 0.4424 $X3$: 0 / 0.2436

Rappel contexte polynomes de chaos - 6

Ex Sobol (entrées U)

$$Y = \prod_{j=1}^{8} \frac{|4X_j - 2| + a_j}{1 + a_j}$$

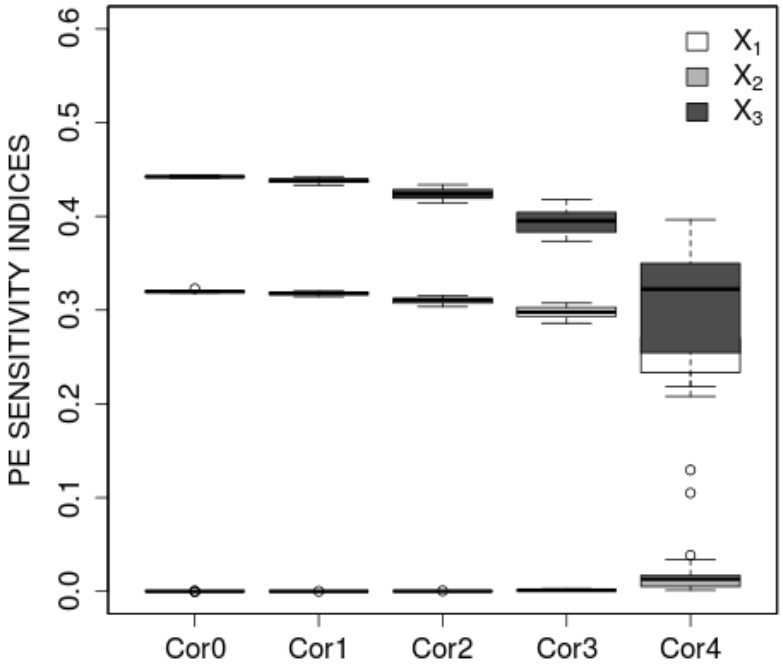


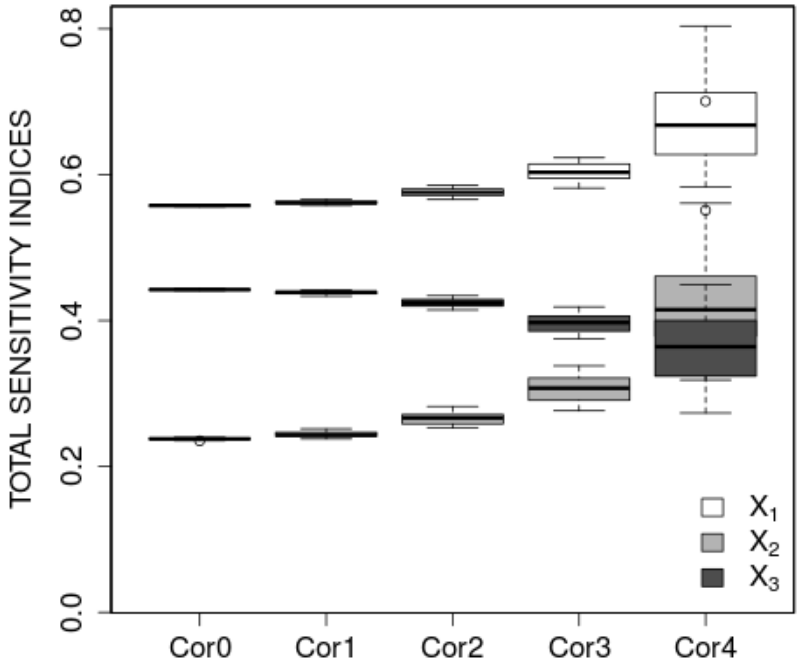
Les indices théoriques :

$X1$: 0.6037/0.6342 $X2$: 0.2683/0.2945 $X3$: 0.0671/0.0756 $X4$: 0.0200/0.0227

Rappel contexte polynomes de chaos - 7

Ex Ishigami: influence de la corrélation entre inputs

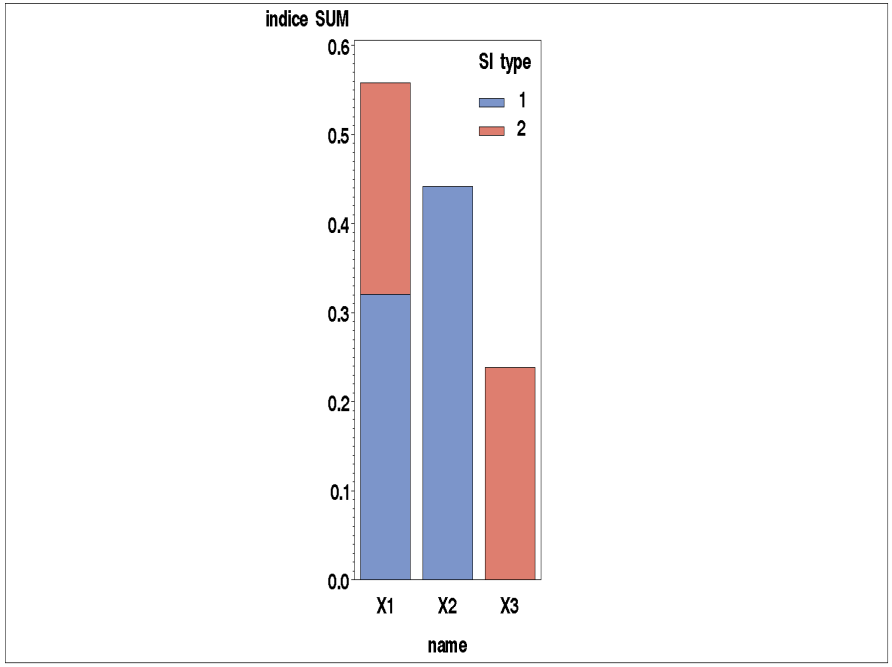


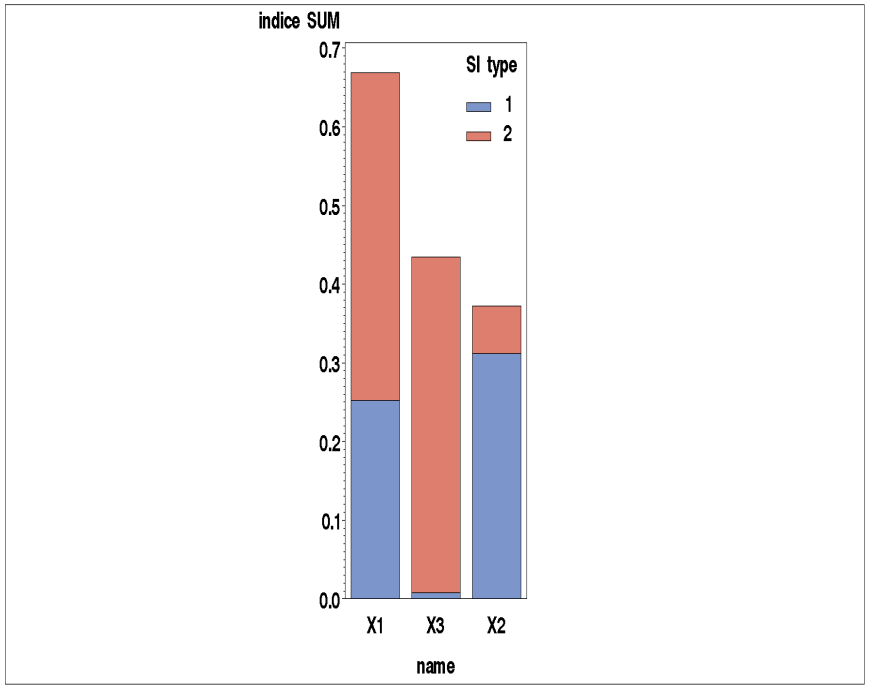


cor=0; cor=[0;0.2]; cor=[0.2;0.4]; cor=[0.4;0.6]; cor=[0.6;0.8]

Rappel contexte polynomes de chaos - 8

Ex Ishigami: entrées U non corrélées et corrélées à 0.80



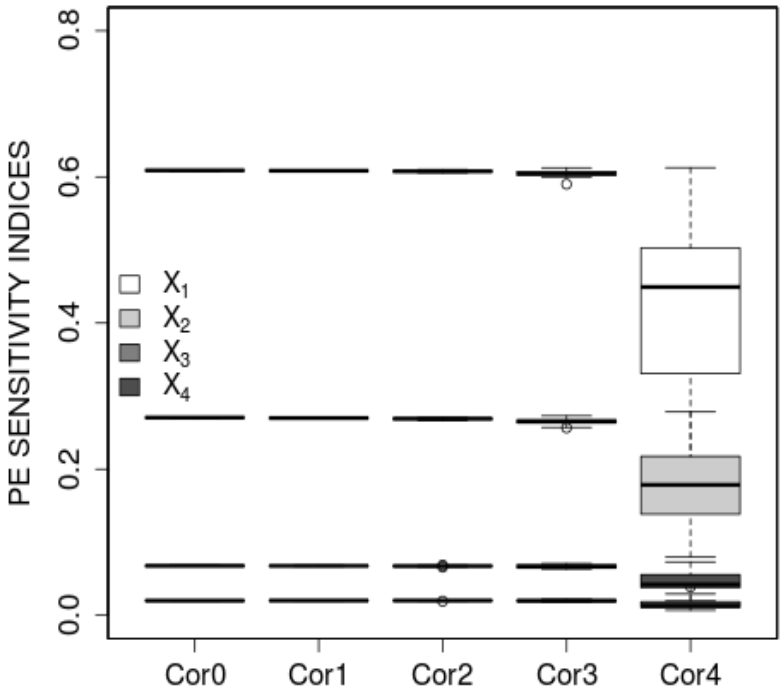


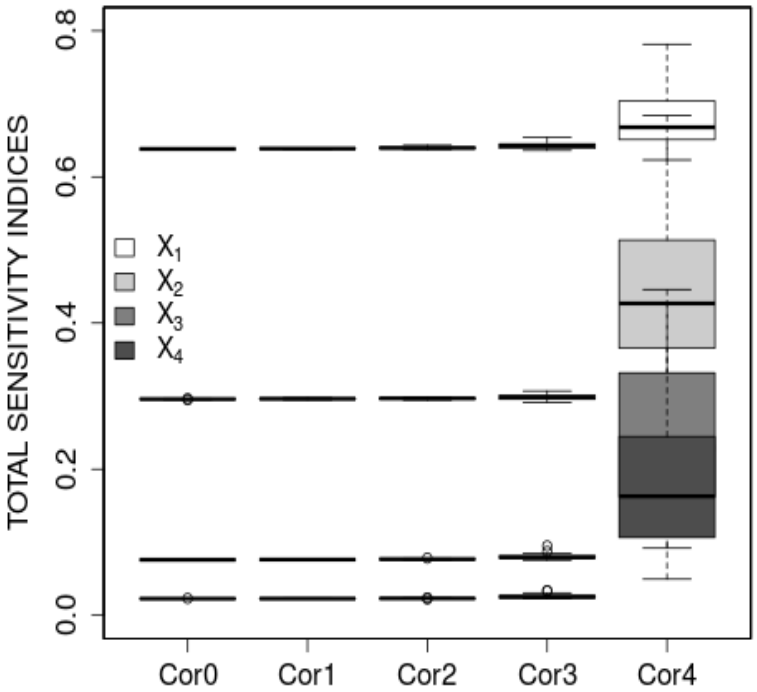
Les indices théoriques (entrées indépendantes) :

$X1 \longrightarrow 0.3138 \ / \ 0.5574 \quad X2 \longrightarrow 0.4424 \ / \ 0.4424 \quad X3 \longrightarrow 0 \ / \ 0.2436$

Rappel contexte polynomes de chaos - 9

Ex Sobol: influence de la corrélation entre inputs

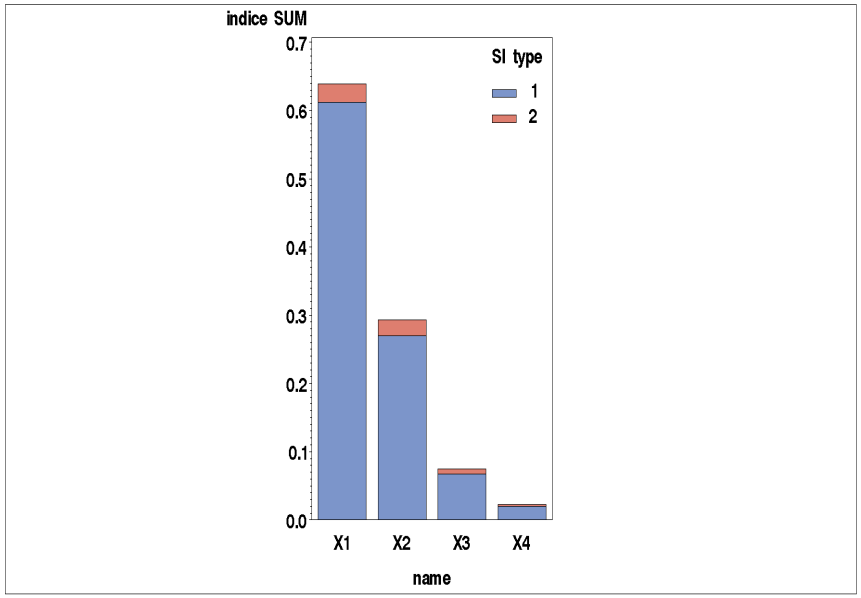


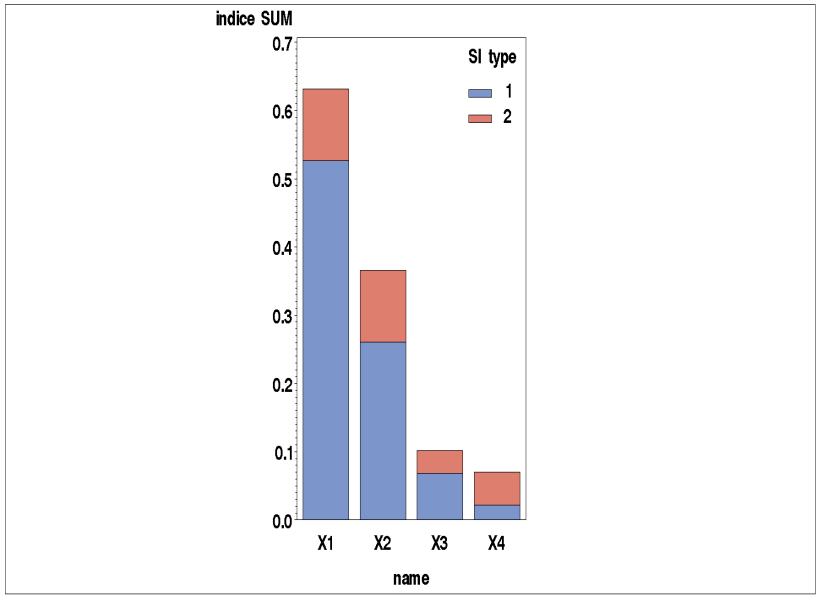


cor=0; cor=[0;0.2]; cor=[0.2;0.4]; cor=[0.4;0.6]; cor=[0.6;0.8]

Rappel contexte polynomes de chaos - 10

Ex Sobol: entrées U non corrélées et corrélées à 0.80





Les indices théoriques (entrées indépendantes) :

$X1$: 0.6037/0.6342 $X2$: 0.2683/0.2945 $X3$: 0.0671/0.0756 $X4$: 0.0200/0.0227

Nouvelle proposition pour des entrées corrélées

Calculer simultanément des nouveaux indices et un métamodèle, en partant d'un polynôme de chaos où les coefficients sont estimés par régression PLS, et non plus par régression OLS, pour des lois quelconques des entrées.

Rappel régression PLS - 1

- Méthode de régression bilinéaire (distribution-free), basée sur la construction de variables latentes, les composantes PLS *th*, qui vise à établir des liens entre un ensemble de L (≥ 1) réponses et un ensemble de P variables explicatives, observés sur les mêmes N objets.
- L'algorithme de PLS est itératif, basé sur l'algo NIPALS (Wold, 1966).
- Pour $L = 1$, on écrira le modèle PLS estimé sur N objets comme

$$\hat{Y}_{h^*} = \Psi(\mathbf{X})\hat{\beta}_{h^*}^{[PLS]}$$

où:

  - $\hat{Y}$ est le $N-$vecteur des valeurs prédites de la réponse par le modèle numérique M,
  - $\Psi(\mathbf{X})$ est la matrice $N \times P$ du modèle de régression construite à partir du plan d'expériences $\mathbf{X}$,
  - le $(P+1)$-vecteur $\hat{\beta}_{h^*}^{[PLS]}$ est l'estimation PLS du vecteur inconnu β des coefficients du modèle de régression, pour le nombre optimal h^* de *th*.

Rappel régression PLS - 2

- Soit $\mathbf{E}_0$ la matrice $\Psi(\mathbf{X})$ centrée et réduite, et F_0 le vecteur réponse Y_N centré-réduit.
- Soit $\mathbf{E}_h$ la matrice résiduelle obtenue suite à la décomposition de $\mathbf{E}_0$ basée sur les h premières t_h.
- Soit F_h le vecteur résiduel obtenu suite à la décomposition de F_0 basée sur les h premières t_h.

OBJECTIFS de l'**étape 1** de l'algorithme:

1. On construit une combinaison linéaire $u_1 = F_0 c_1$, et une combinaison linéaire $t_1 = \mathbf{E}_0 w_1$, par maximisation de la covariance entre t_1 et u_1 sous les contraintes $\|w_1\|_2 = \|c_1\|_2 = 1$.
2. Puis par régressions (simples !) on obtient les vecteurs de coefficients p_1 et r_1 qui permettent les décompositions (déflations)

$$\mathbf{E}_0 = t_1 p_1^T + \mathbf{E}_1$$
$$F_0 = t_1 r_1^T + F_1$$

Rappel régression PLS - 3

- **Etape 2**: on remplace $\mathbf{E}_0$ et F_0 par $\mathbf{E}_1$ et F_1, et on recommence comme à l'étape 1.
- Pour h^*étapes on obtient (avec des th toutes $\perp$ entre elles)

$$F_0 = t_1 r_1^T + \ldots + t_h r_{h^*}^T + F_{h^*}$$

- On peut écrire th sous sa forme développée:

$$t_h = \left[\sum_{j=1}^{P} cov^2\left(F_{h-1}, E_{h-1,j}\right)\right]^{-1/2} \sum_{j=1}^{P} \left[cov\left(F_{h-1}, E_{h-1,j}\right)\right] E_{h-1,j}$$

- On observe que le terme $cov\left(F_{h-1}, E_{h-1,j}\right)$ est une covariance entre les résidus F_{h-1} et $E_{h-1,j}$
$\Longrightarrow$ c'est donc une **covariance partielle** par definition: c'est la raison pour laquelle la multicolinearité entre les variables explicatives X_j est efficacement prise en comte à chaque étape.

Rappel régression PLS - 4

On déduit les équations importantes:

$$\|F_0\|^2 = var(F_0) = r_1^2 \|t_1\|^2 + \ldots + r_{h^*}^2 \|t_{h^*}\|^2 + \|F_{h^*}\|^2$$

$$\hat{F}_0 = \sum_{h=1}^{h^*} t_h r_h^T \quad \Longrightarrow \text{var}\left(\hat{F}_0\right) = \sum_{h=1}^{h^*} r_h^2 var(t_h)$$

- Si $\hat{F}_0$ est un bonne approximation de F_0 alors:
  - on dispose d'une bonne approximation de la décomposition **orthogonale** de la variance de la réponse (réduite) via les t_h,
  - les r_h^2 sont des indices de sensibilité
- En développant les t_h on a aussi:

$$\hat{Y}_{h^*} = \hat{\beta}_{h^*,0}^{[PLS]} + \hat{\beta}_{h^*,1}^{[PLS]} \Psi_1(\mathbf{X}) + \ldots + \hat{\beta}_{h^*,P}^{[PLS]} \Psi_P(\mathbf{X})$$

Rappel régression PLS - 5

La validation d'une composante *th* du modèle PLS peut se baser sur l'indice spécifique suivant:

$$Q_h^2 = 1 - \frac{\text{PRESS}_h}{\text{RSS}_{h-1}}$$

où:

$$\text{RSS}_{h-1} = \sum_{i=1}^{N} \left(y_i - \hat{y}_i^{[h-1]} \right)^2 = \sum_{i=1}^{N} e_{h-1,i}^2 \quad \text{et} \quad \text{PRESS}_h = \sum_{i=1}^{N} \left(\frac{e_{hi}}{1 - \Gamma_{h,ii}} \right)^2$$

où: $e_{hi} = y_i - \hat{y}_i^{[h]}$, et $\Gamma_{h,ii}$ est l'élément diagonal de $\Gamma_h = T(T^T T)^{-1} T^T$ où $T = [t_1 || t_2 || ... || t_h]$.

Rappel régression PLS - 6

Règle d'arrêt dans le contexte de cette méthode :

- On accepte la composante courante th tant que $Q_h^2 > 0$ et que $|Q_h^2 - Q_{h-1}^2| > 10^{-4}$
- Soit h^* la dernière composante satisfaisant à ces contraintes.
- La qualité finale du modèle est alors synthétisée dans le critère:

$$Q_{cum}^2(h^*) = 1 - \prod_{h=1}^{h^*} \left(1 - Q_h^2\right)$$

Nouveaux indices de sensibilité

- Ils sont obtenus en remplaçant $\hat{\beta}_k^{[OLS]}$ par $\hat{\beta}_{h^*,k}^{[PLS]}$ dans les IMSI_k.
- Deux propriétés fondamentales:
  - **P1:** si $L = 1$ et $\Psi(\mathbf{X})$ orthogonale alors $h^* = 1$ et on a $\hat{\beta}_1^{[PLS]} = \hat{\beta}^{[OLS]}$ et donc:

$$\begin{aligned} \text{PC}_d\text{-PESI(PLS)} &= \text{PC}_d\text{-PESI(OLS)} \\ \text{PC}_d\text{-TSI(PLS)} &= \text{PC}_d\text{-TSI(OLS)} \end{aligned}$$

  - **P2**: Si $\Psi\left(\tilde{\mathbf{X}}\right)$ est non orthogonale (où $\tilde{\mathbf{X}}$ est une matrice de modèle corrélée), de rang plein: $A = P + 1$, ou non: $A < P + 1$, alors

$$\hat{\beta}_h^{[PLS]} \underset{h \longrightarrow A}{\longrightarrow} \hat{\beta}^{[OLS]}$$

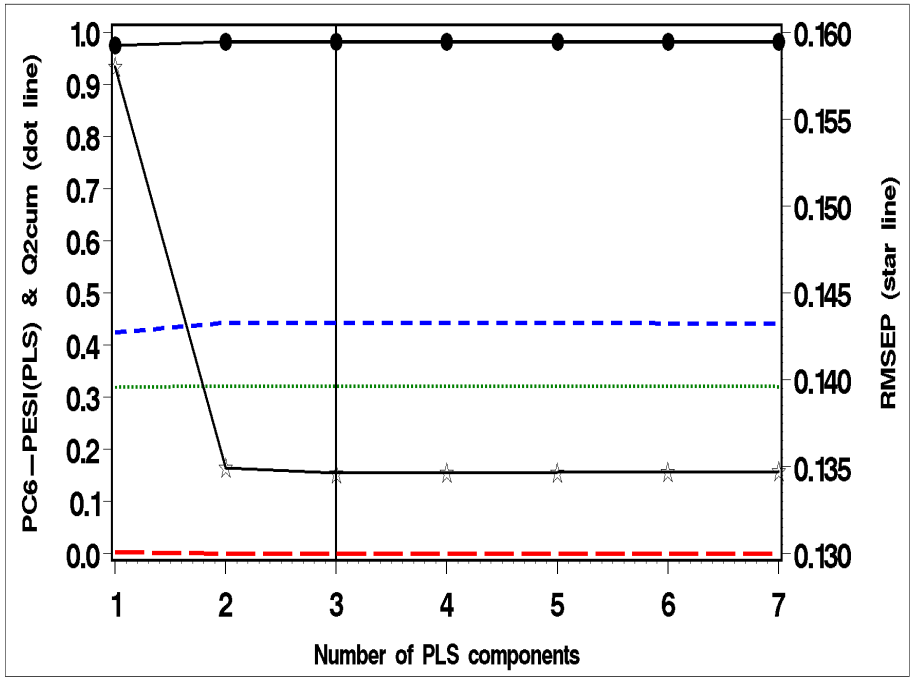
et donc:

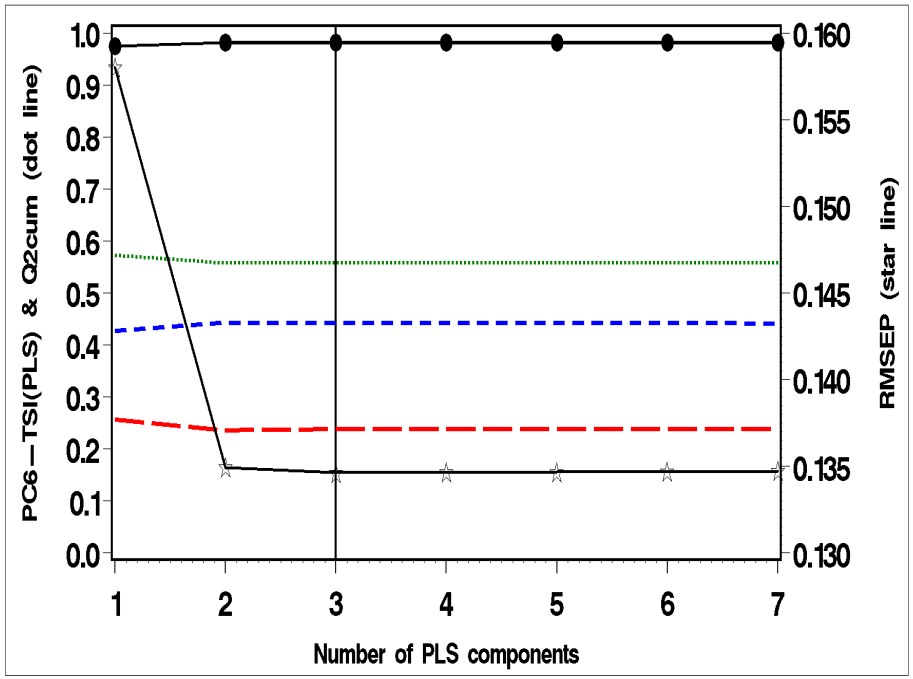
$$\begin{aligned} \text{PC}_d\text{-PESI(PLS)} &\longrightarrow \text{PC}_d\text{-PESI(OLS)} \\ \text{PC}_d\text{-TSI(PLS)} &\longrightarrow \text{PC}_d\text{-TSI(OLS)} \end{aligned}$$

Avantages de la méthode

- Les PC_d-PESI(PLS) et les PC_d-TSI(PLS) traduisent bien les influences des inputs grâce à l'efficacité de la gestion de la multicolinéarité de la régression PLS, contrairement à la régression OLS qui gère mal les fortes corrélations: les $\hat{\beta}^{OLS}$ ne sont plus interprétables, et en conséquence les PC_d-PESI(OLS) et les PC_d-TSI(OLS) non plus.
- Le nombre N de runs peut être inférieur au nombres de monôme dans le polynôme (avantage majeur quand les runs sont "chers").
- Pas d'inversion de matrice avec l'algorithme de PLS et donc même pour des monômes fortement corrélés (situation fréquente pour des entrées elles-mêmes seulement modérément corrélées), et a fortiori pour des relations fonctionnelles entre certaines inputs, le calcul des indices reste possible (contrairement à la régression OLS qui ne fonctionne plus car la matrice d'information devient quasi-singulière ou singulière).

Ex Ishigami: Entrées non corrélées

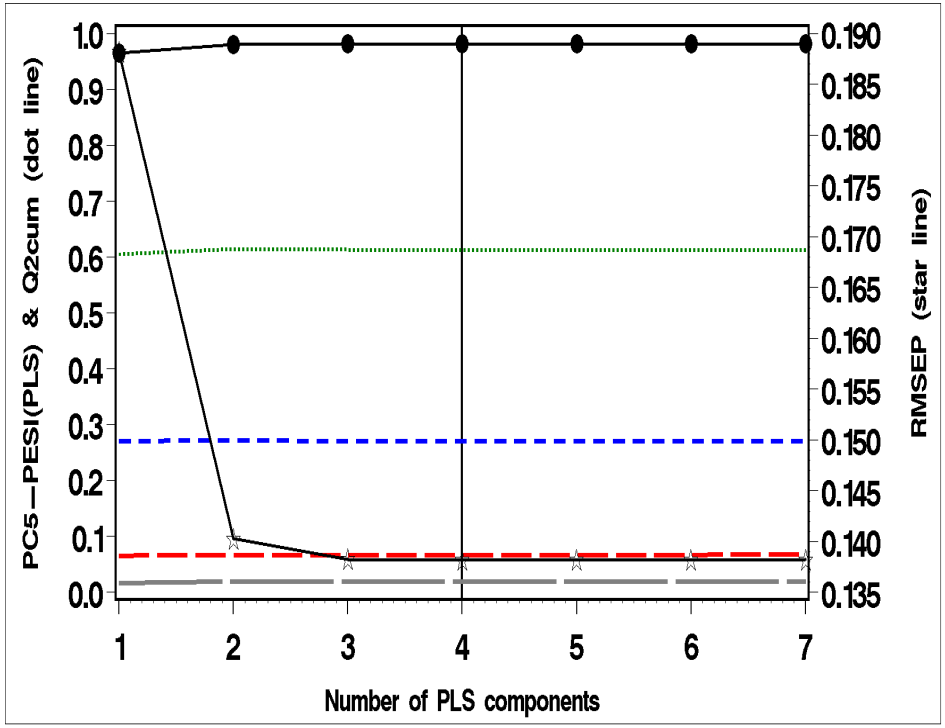


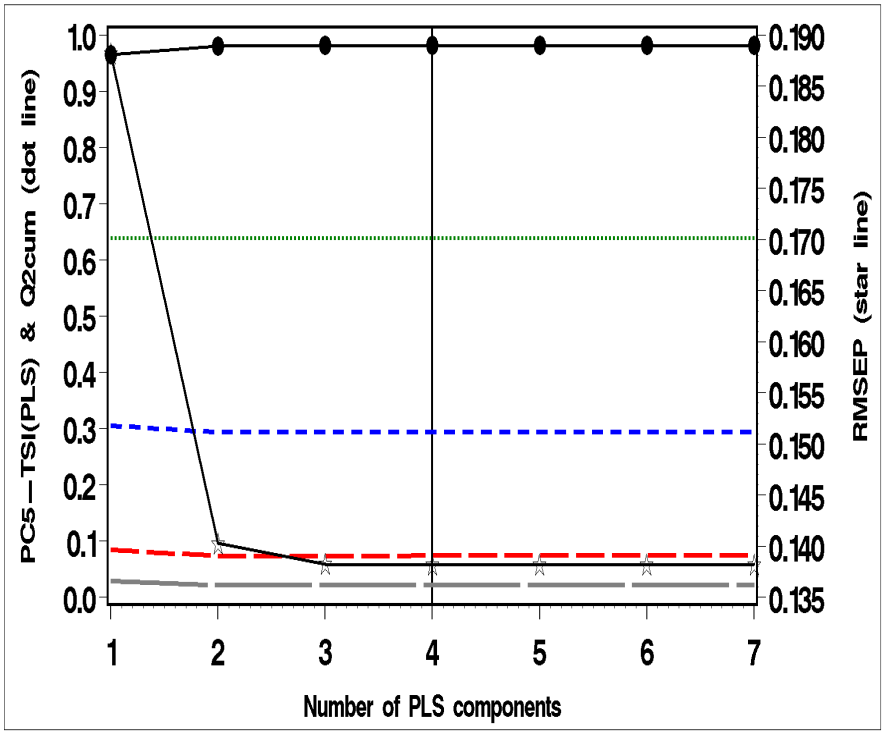


Les indices théoriques (entrées indépendantes):

$X1 \longrightarrow 0.3138 \ / \ 0.5574 \quad X2 \longrightarrow 0.4424 \ / \ 0.4424 \quad X3 \longrightarrow 0 \ / \ 0.2436$

Ex Sobol: Entrées non corrélées

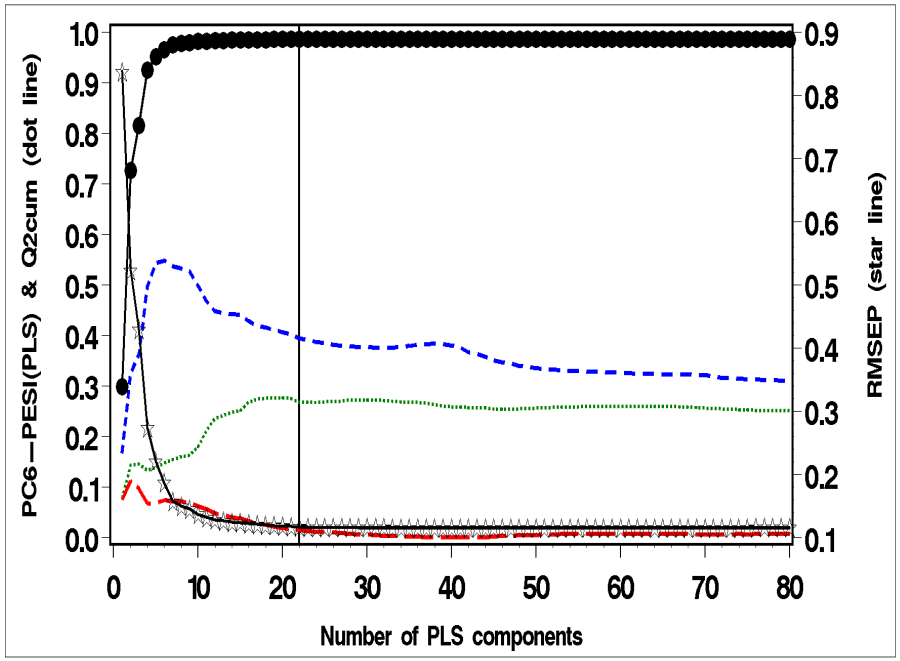


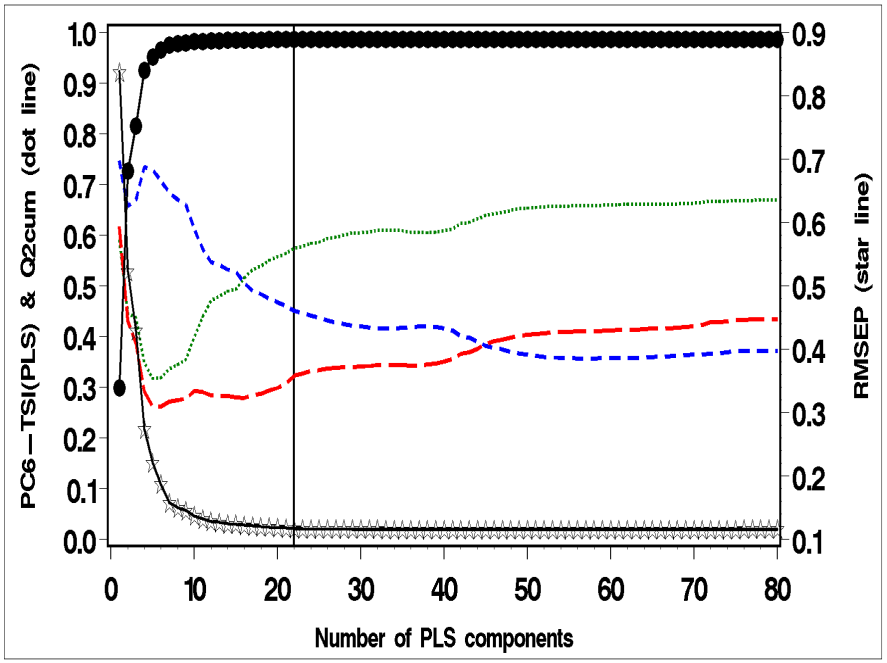


Les indices théoriques (entrées indépendantes) :

$X1$: 0.6037/0.6342 $X2$: 0.2683/0.2945 $X3$: 0.0671/0.0756 $X4$: 0.0200/0.0227

Ex Ishigami: Entrées corrélées à 0.8

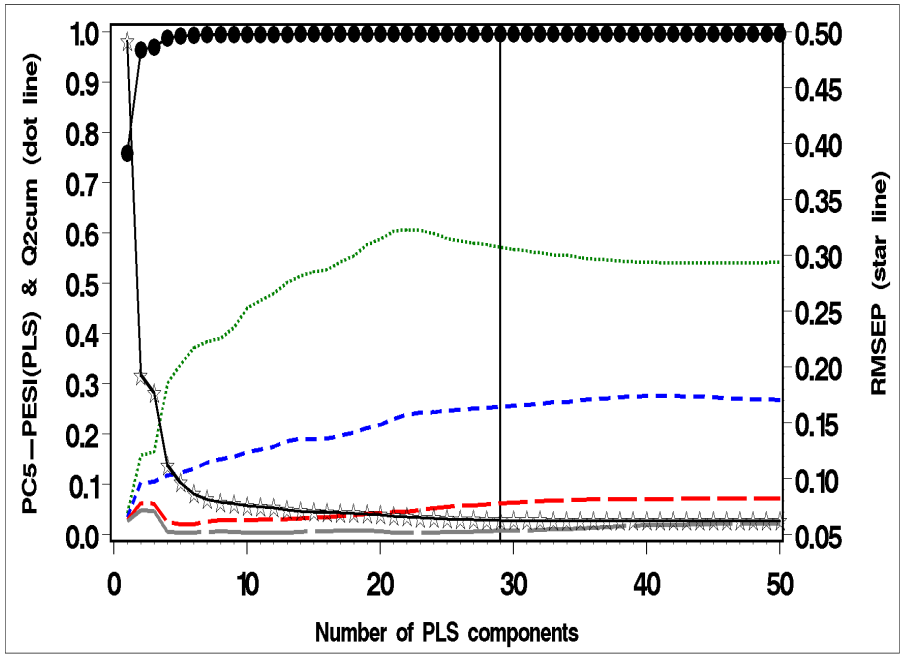


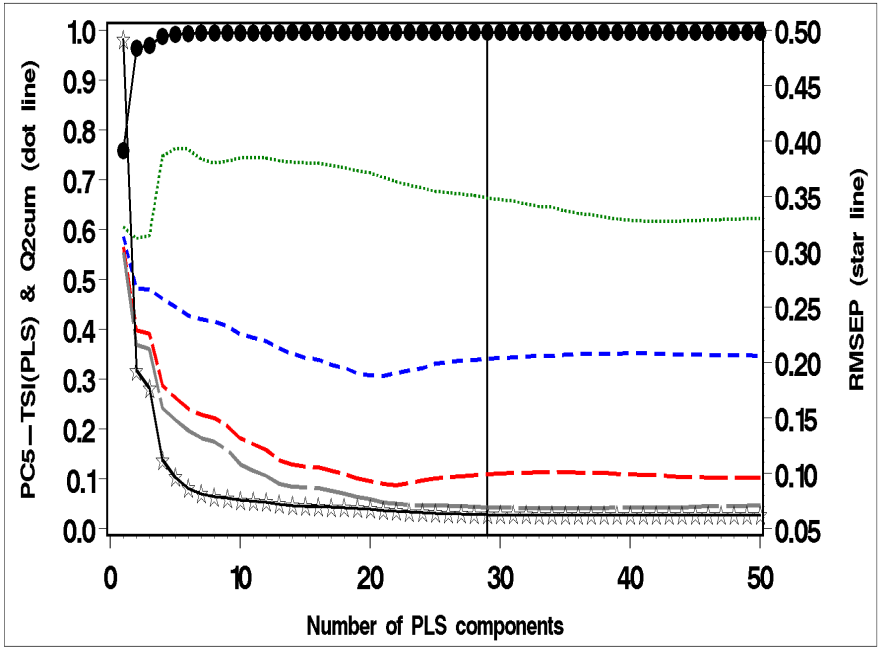


Les indices théoriques (entrées indépendantes) :

$X1 \longrightarrow 0.3138 \ / \ 0.5574 \quad X2 \longrightarrow 0.4424 \ / \ 0.4424 \quad X3 \longrightarrow 0 \ / \ 0.2436$

Ex Sobol: Entrées corrélées à 0.8





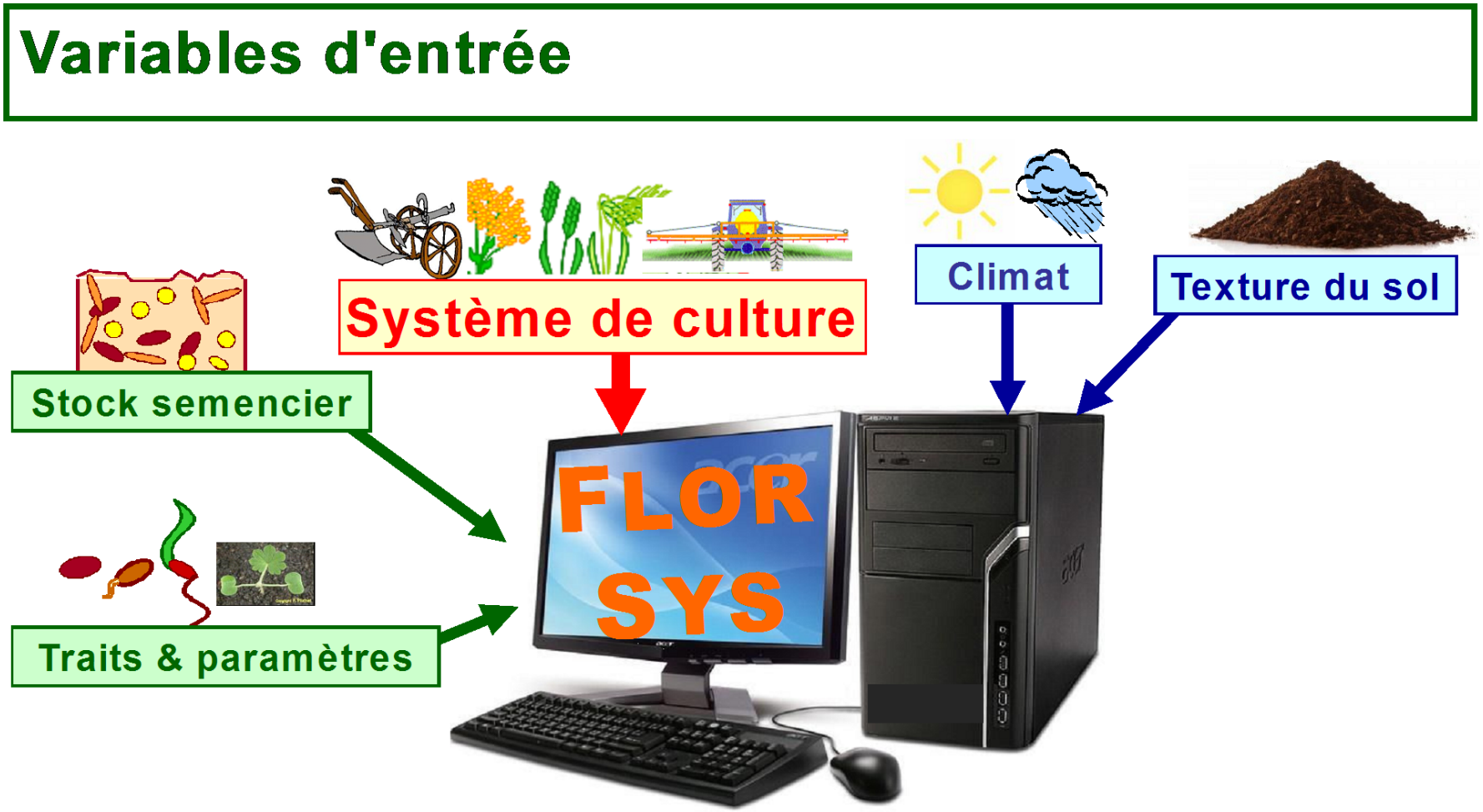
Les indices théoriques (entrées indépendantes) :

$X1$: 0.6037/0.6342 $X2$: 0.2683/0.2945 $X3$: 0.0671/0.0756 $X4$: 0.0200/0.0227

FLORSYS : c'est quoi ?

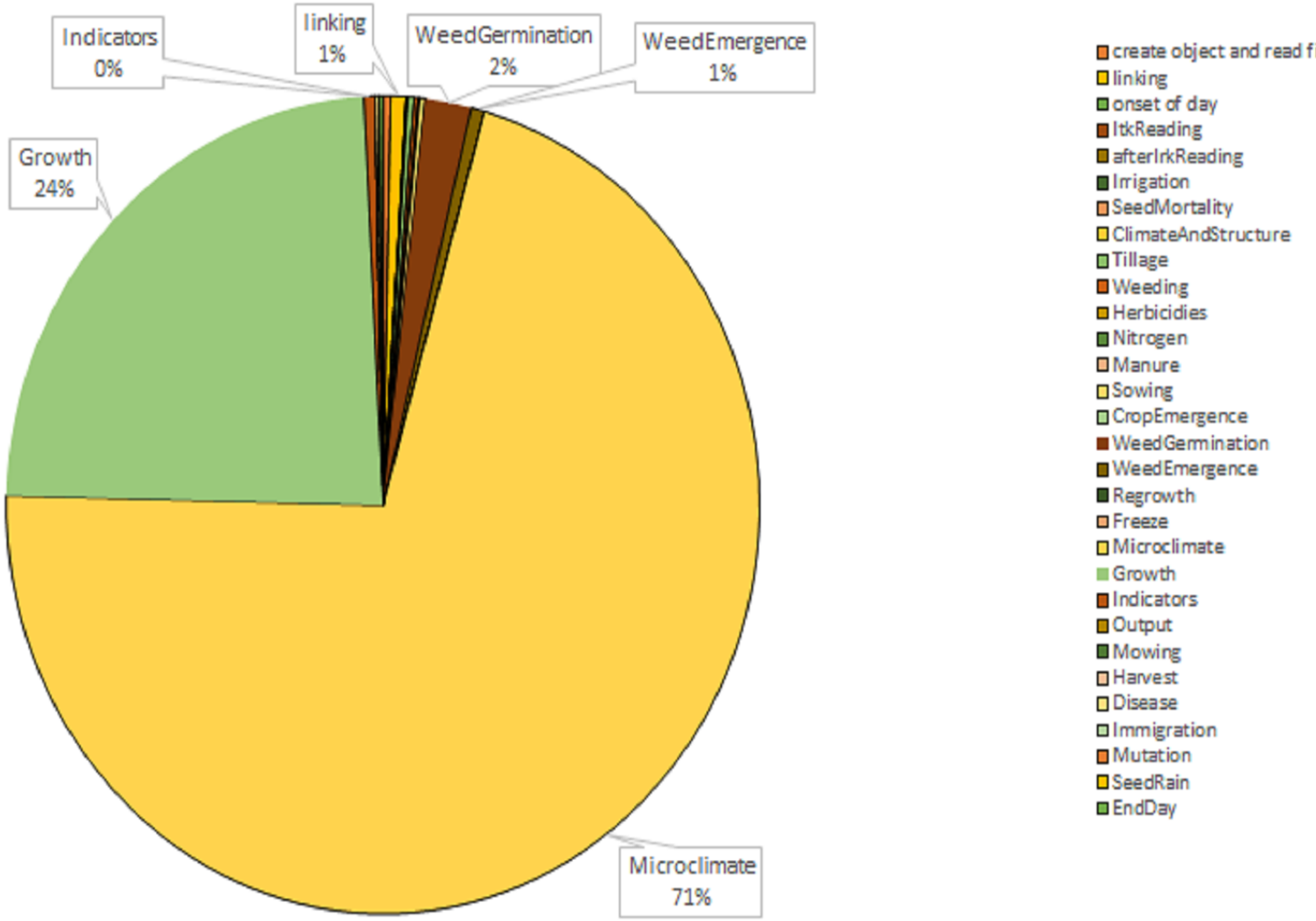
- C'est un modèle numérique qui sert à la recherche en agronomie, codé en langage C/C++ et formé de nombreux modules.
- Il permet:
  - des expérimentations virtuelles,
  - l'évaluation de systèmes de culture actuels et prospectifs,
  - de concevoir de nouveaux systèmes de culture,
  - d'étudier l'optimisation de l'utilisation des adventices en vue de la diminution des doses d'herbicides (Agro-Ecologie)

FLORSYS : un modèle mécaniste complexe



Colbach et al (2006, 2010) Eur J Agron, Colbach et al (2007) Ecol Mod; Gardarin et al. (2012) Ecol Mod; Munier-Jolain et al. (2013) Ecol Mod, Colbach et al. (submitted) Eur J Agron

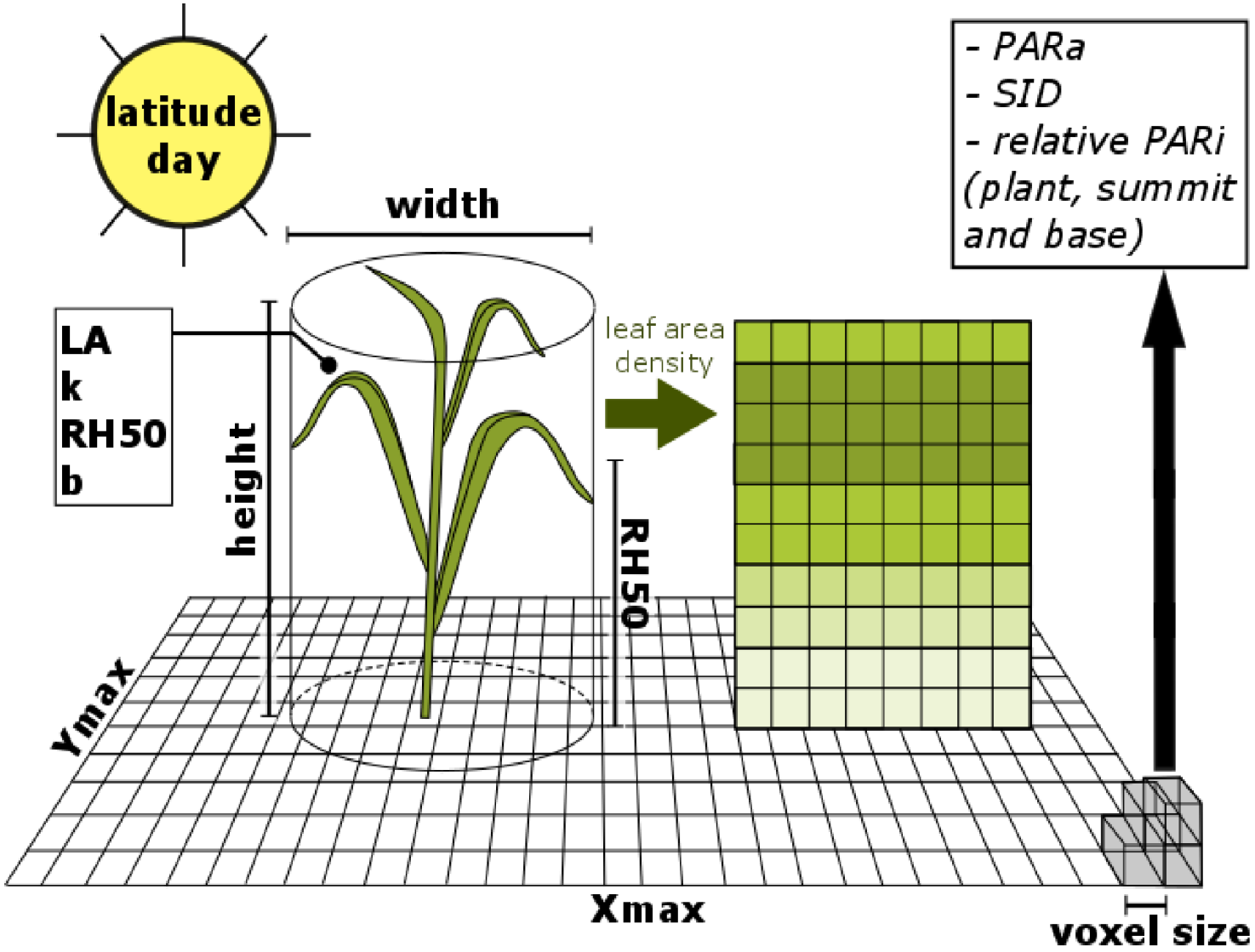
FLORSYS : répartition temps calcul entre les modules



FLORSYS : premier objectif de la thèse de F. Colas

- Une simulation sur 30 années avec 10 répétitions climatiques et pas-de-temps journalier ===> 30 heures environ
- Le module "microclimat" : 20 heures de calcul environ
- Le premier objectif de la thèse de F. Colas est de faire l'analyse de sensibilité de FLORSYS (plusieurs centaines d'entrées, nombreuses sorties):
  ====> Il faut d'abord accélérer le module "microclimat"
- Le moyen choisi: la métamodélisation du module avec analyse de sensibilité conjointe de ses entrées:
  - pour une plante seule isolée des autres plantes (11 entrées),
  - pour une plante avec son voisinage "couvert" (15 entrées), environ 10000 plantes sur la parcelle.

FLORSYS : le module "microclimat"



Module "microclimat" - plante seule

11 entrées continues (de loi uniformes) prises en compte, 6 corrélées
====> matrice de corrélation (après Iman & Conover):

	Height	Width	LA	k	RH50	b
Height	1	0.58	0.19	−0.19	0.20	0.29
Width	0.58	1	0.19	0	0.19	0.30
LA	0.19	0.19	1	0	0	0
k	−0.19	0	0	1	0.20	−0.11
RH50	0.20	0.19	0	0.20	1	0.09
b	0.29	0.30	0	−0.11	0.09	1

Module "microclimat" : objectifs

1. Calcul conjoint de métamodèles et d'indices de sensibilité des 11 entrées sur 5 sorties (liées à l'incidence du rayonnement solaire sur les feuilles), contexte "plante seule"
2. Mêmes calculs pour "plante seule + son voisinage couvert" (15 inputs)

Sous les contraintes:

- Une prédiction d'un métamodèle < 0.2 seconde ($\simeq$ temps CPU actuel du module microclimat pour un run).
- Prendre en compte les corrélations entre inputs.

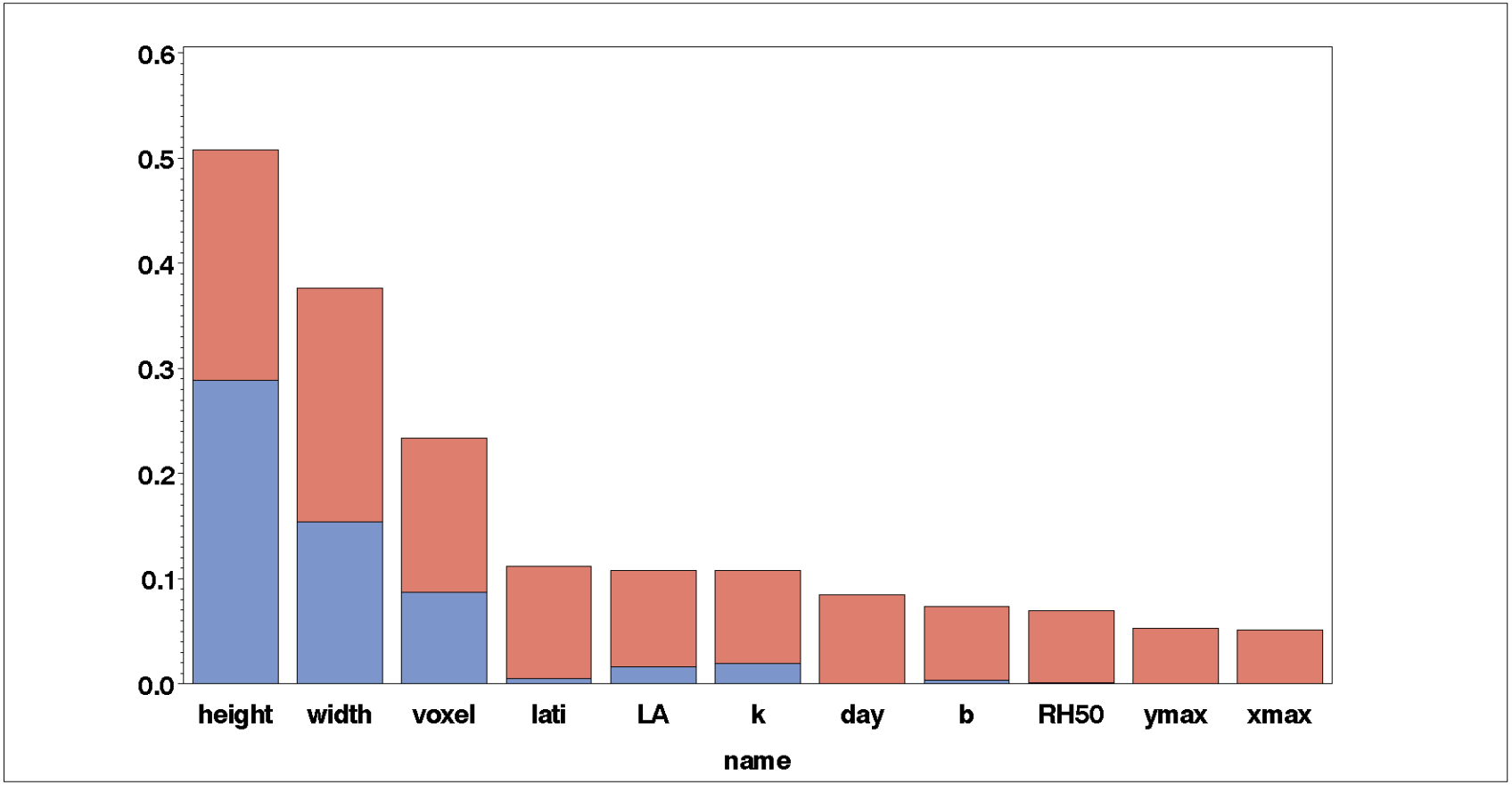
Module "microclimat" : résultats (1a)

M#	d	Monomial number	h^*	$Q^2_{cum}(h^*)$	R^2	$RMSEP_{h^*}$	CTB#
1	1	$P = 11$	3	0.56	0.56	0.662	3"
2	2	$P = 77$	6	0.80	0.80	0.450	24"
3	3	$P = 363$	8	0.87	0.87	0.358	2'
4	4	$P = 1364$	12	0.91	0.92	0.290	10'
5	5	$P' = 2000$	15	0.93	0.94	0.250	20'
6	5	$P' = 3000$	18	0.94	0.95	0.224	1h

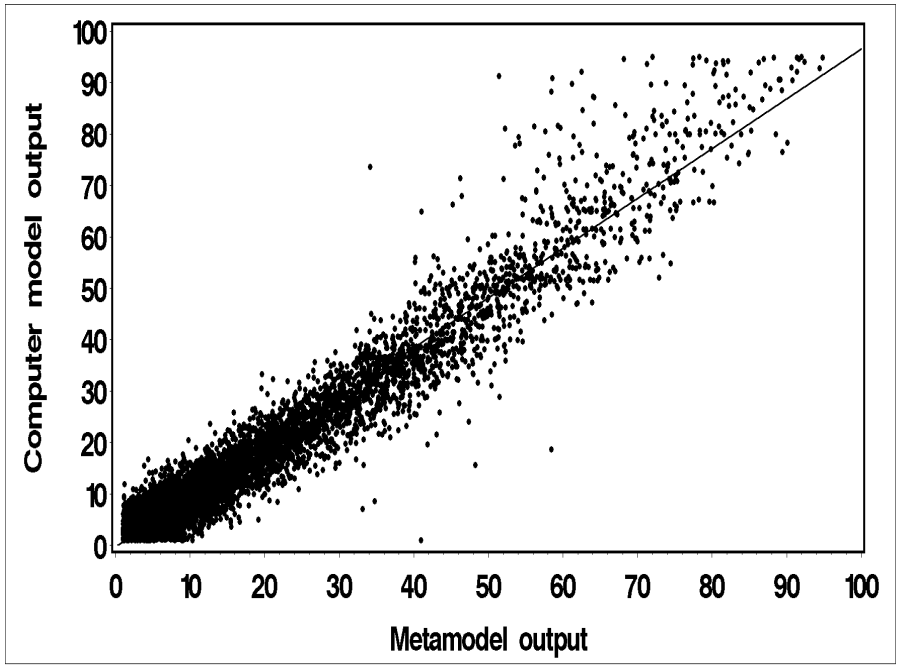
Module "microclimat" : résultats (1b)

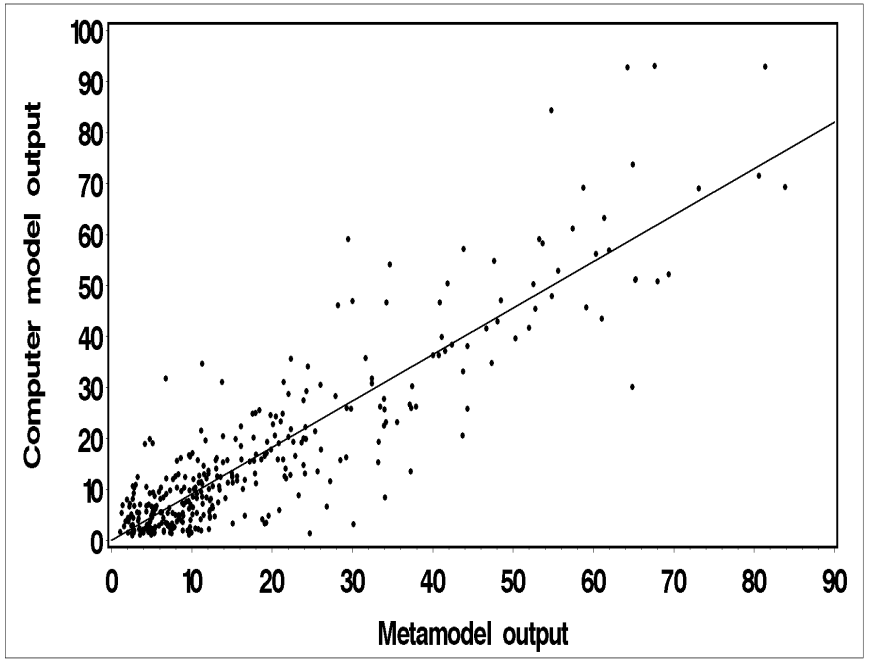
M#	d	Monomial number	h^*	$Q^2_{cum}(h^*)$	R^2	$RMSEP_{h^*}$	CTB#
7	5	$P = 4367$	22	0.96	0.96	0.195	4h
8	6	$P' = 2000$	17	0.94	0.95	0.241	28'
9	6	$P' = 3000$	20	0.95	0.96	0.212	1h 18'
10	6	$P' = 4000$	24	0.96	0.97	0.186	3h 30'
11	7	$P' = 2000$	19	0.93	0.94	0.246	43'
12	7	$P' = 3000$	22	0.95	0.95	0.215	1h 42'
13	7	$P' = 4000$	28	0.96	0.97	0.186	5h

Module "microclimat" : résultats (2) - modèle M#7



Module "microclimat" : résultats (3) - reconstitution et prédiction





$R^2 = 0.960 \longrightarrow R^2 = 0.789$

Module "microclimat" : conclusions

- On dispose des indices de sensibilité polynomiaux et totaux pour les entrées considérées, obtenus à partir des métamodèles "complets" :
  ===> plusieurs entrées à faible indice total
  ===> simplification ultérieure en fixant ces entrées à des valeurs moyennes ou de référence
  ===> nouveaux métamodèles plus "légers"
- Temps de calcul d'une prédiction par M#7 $\simeq$ la moitié du temps de calcul par le modèle numérique "micro-climat", soit 0.1 seconde.
- Réduction non négligeable si on considère que 10000 plantes sont à simuler.
- Essais en cours sur FlorSys global avec 4 métamodèles parcimonieux ($<$ 30 monômes) et pour un voxel de 1 cm de côté: on arrive à diviser le temps total par 10 selon les situations.
- ASG de FLORSYS envisageable

Productions

- Deux packages R sur le CRAN sont disponibles (GPL license):
  - https://CRAN.R-project.org/package=plspolychaos
  - https://CRAN.R-project.org/package=polychaosbasis
- Une publication en révision (EMS)

Améliorations (1)

- Trop de monômes ?
  - limiter le nombre de monômes a priori
  - limiter le nombre de monômes a posteriori
- Limiter le nombre de runs ?

Améliorations (2)

Limiter le nombre de monômes a priori :

- On sélectionne les P' meilleurs monômes parmi les P monômes du polynôme complet de degré d, au sens du R^2 de chacune des P régressions simples réalisées lors d'une étape préliminaire.
- Exemple Ishigami:

$P' =$13 ($R^2 = 0.9997$) et $P = 164$ pour $d = 8$ ($R^2 = 0.9977$):

$$\hat{Y} = f(X, X2, X3, X_2^4, X_2^6, X_1X_3^2, X_1^3, X_1^3X_3^2, X_2^2, X_1X_3^4, X_2^8, X_1^3X_3^4, X_1^5)$$

Améliorations (3)

Limiter le nombre de monômes a posteriori :

- LASSO + PLS
- Exemple "microclimat": à partir de M#7$\longrightarrow$ 25 monômes seulement:

$$\hat{Y} = f(X_j, j = 1, , ..., 11, \text{ et 14 autres monômes interactifs})$$

- Q^2_{cum} ($d = 5$, 4367 mono) $= 0.95$, CPU d'une prédiction $= 0.1$ seconde.
- Q^2_{cum} ($d = 5$, 25 mono) $= 0.81$, CPU d'une prédiction < 0.01 seconde.

... mais en général les indices sont modifiés.

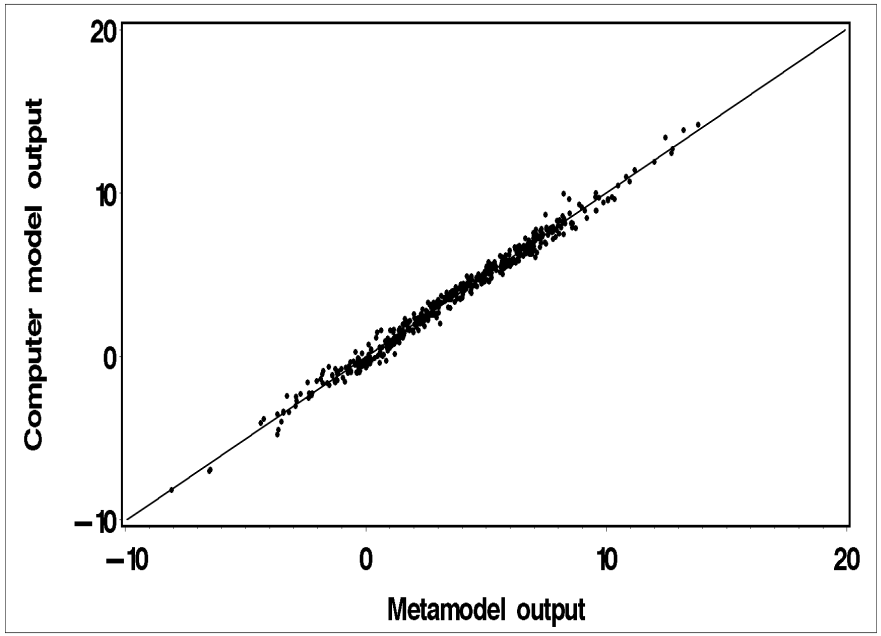
Améliorations (4)

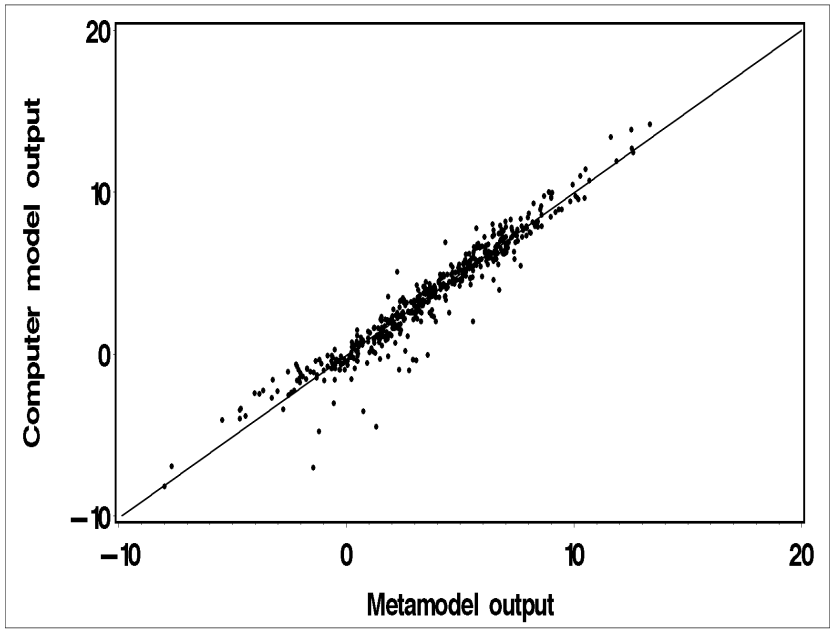
Limiter le nombre de runs (chers !) $\longrightarrow$ une approche (séquentielle) empirique:

1. On tire aléatoirement un plan d'expériences de petite taille par exemple $(n = 100)$ dans un grand SFD.
2. A partir des n runs $y_i = M(\mathbf{x}_i)$ et des Kn runs issus des K plans précédents, on calcule par régression PLS le métamodèle $\hat{Y} = M_d(\mathbf{x}_i)$, et les indices, en augmentant progressivement d jusqu'à $Q^2_{cum} \geq$ seuil. Si échec on repart en 1.
3. On cherche un métamodèle parcimonieux de $\hat{Y} = M_{d^*}(\Psi(\mathbf{X}))$

Améliorations (5)

Limiter le nombre de runs (chers !): exemple Ishigami:





Apprentissage $N = 9500 \longrightarrow R^2$ sur le test $= 0.987$
Apprentissage $N = 100 \longrightarrow R^2$ sur le test $= 0.930$

Question ouverte

Limiter le nombre total de runs avec une approche optimale sur la structure du plan d'expériences séquentiel X de taille $n << P$ et éventuellement sous la contrainte supplémentaire $N = Sn < P$ pour S plans successifs.

Perspectives

- Analyse de sensibilité dynamique (évolution des indices en fonction du temps) $\longrightarrow$ package R "multisensi" conçu par l'INRA (Unité MaIAGE - Jouy-en-Josas).
- Analyse de sensibilité pour des sorties fonctionnelles (courbes, spectres, ...) à partir des composantes de l'ACP$\longrightarrow$ package R "multisensi".
- Analyse de sensibilité généralisée pour un mélange d'entrées continues et d'entrées qualitatives, et des sorties univariées ou multivariées $\longrightarrow$ méthode SIVIP en cours de publication.

MERCI DE VOTRE ATTENTION !!!