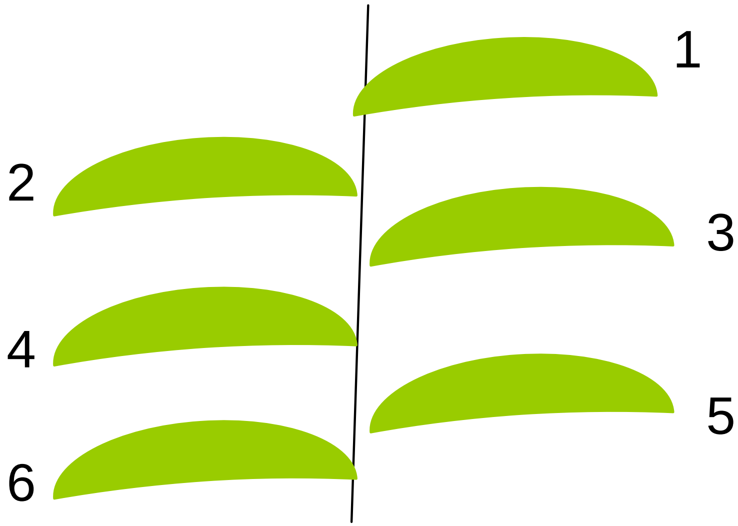
Application de méthodes statistiques aux modèles dynamiques pour l'agronomie et l'élevage

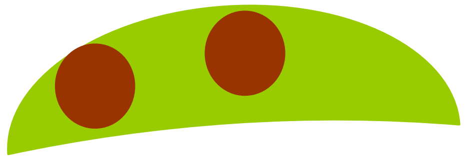
cas d’étude Septolis

Modèle pour septoriose du blé

Daniel Wallach (INRA)

Le système





Y(t,l)=fraction de la feuille l infectée jour t

(ici Y(t,l)~0.3)

Le modèle

- David Gouache, Arvalis Institut du végétal
- D’après Audsley, Milne, Paveley, 2005, Ann Appli Bot

Équation de base

- Fraction de la feuille l infectée, jour t

$$Y(t,l)=\sum_{i=1}^{t} I(i,l)\,y\left[T(t)-T(i)\right]$$

$I(i,l)=$ nombre d'infections jours précédents

$y\left[T(t)-T(i)\right]=$ taille d'une infection du jour i

Chaque infection a une croissance logistique

Nombre d’infections chaque jour

- Calculer production de spores
- Spores qui arrivent sur chaque feuille
- Spores qui donnent une infection

Production de spores par des infections

- Ça dépend de la surface infectée

$$U(t,l) = 1 - e^{-a\left[Y(t,l) + I_s \delta(l-6)\right]}$$

Spores qui arrivent à la feuille l

$$F(t,l) = U_g e^{-kh_g(t)} + \sum_{j=l+1}^{6} U(t,j) e^{-kh_j(t)} + \rho U(t,l)$$

= spores du sol

+ spores des autres feuilles

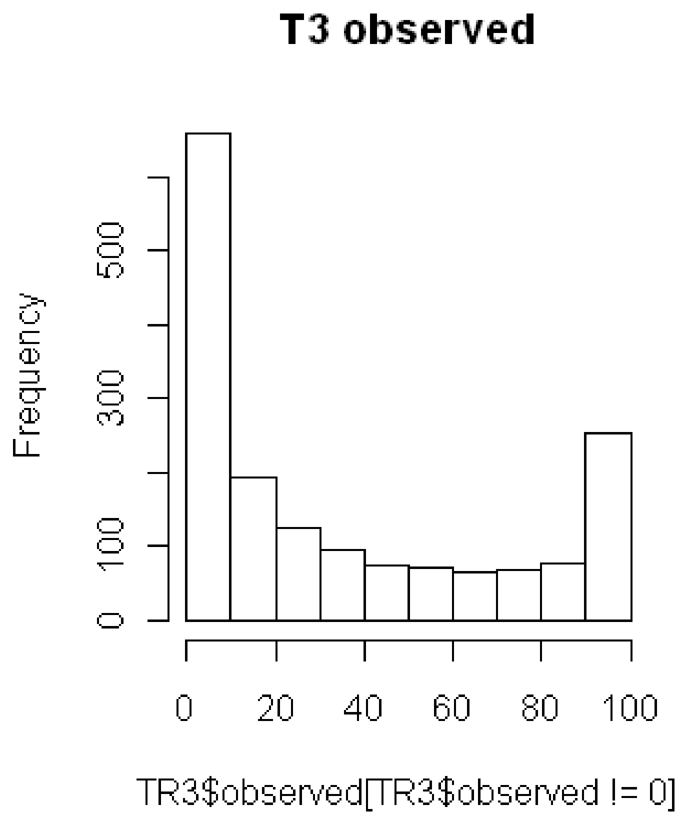
+ auto-infection

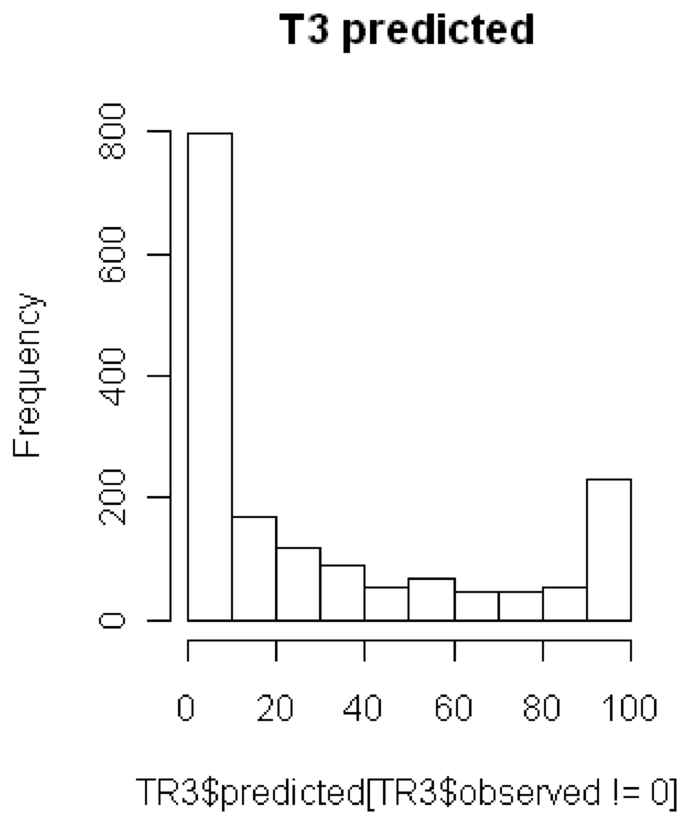
Nombre d'infections

$$I(t,l) = F(t,l)\, f_{pluie}\, f_{niveau\ maladie}\, f_{fraction\ visible}\, f_{fraction\ saine}$$

Les données

- 2001 données 157 sites





Les transformations considérées (données et modèle)

data set	data transformation
T1	$y_{sfi} = \sin^{-1}\left(\sqrt{d_{sfi}}\right)$
T2	$y_{sf1} = \sin^{-1}\left(\sqrt{d_{sf1}}\right)$ $y_{sfi} = \sin^{-1}\left(\sqrt{d_{sfi} - d_{sf(i-1)}}\right) \quad i > 1$
T3	$y_{sfi} = d_{sfi}$
T4	$y_{sf1} = d_{sf1}$ $y_{sfi} = d_{sfi} - d_{sf(i-1)} \quad i > 1$

Les paramètres

- Estimer 3 paramètres par moindres carrés
- Autres paramètres de la littérature

$$\hat{\theta} = \arg\min_{\theta} \left[\sum \left(Y - f(X;\theta,\theta_{lit}) \right)^2 \right]$$

	θ_1 (coeff.mult)	θ_2 (echelle)	θ_3 (prod.inoc)
T1	7.05	13.2	18.6
T2	4.85	11.5	15.9
T3	7.89	14.8	16.9
T4	4.70	11.3	29.5

Objectif

- Estimer la qualité de prédiction
  - Variable: fraction surface infectée, toutes feuilles confondues, toutes dates confondues
    - C'est ce que l'on mesure
  - Population cible: Parcelles agriculteurs de blé en France
  - Critère: erreur de prédiction moyenne, intervalles de confiance

Qualité d'ajustement

$$MSE = \sum (y_i - f(X_i;\hat{\theta}_{MV}))^2 / N = \hat{\sigma}^2_{MV}$$

$$MSE = biais^2 + SDSD + LCS$$

$$biais^2 = \left[(1/N) \sum (y_i - f(X_i;\hat{\theta}_{MV}) \right]^2$$

$$SDSD = (\sigma_y - \sigma_{f(X;\hat{\theta}_{MV})})^2$$

$$LCS = 2\sigma_y \sigma_{f(X;\hat{\theta}_{MV})} r$$

$$EF = 1 - \frac{\sum \left(y_i - f(X_i;\hat{\theta}_{MV} \right)^2}{\sum \left(y_i - \bar{y} \right)^2}$$

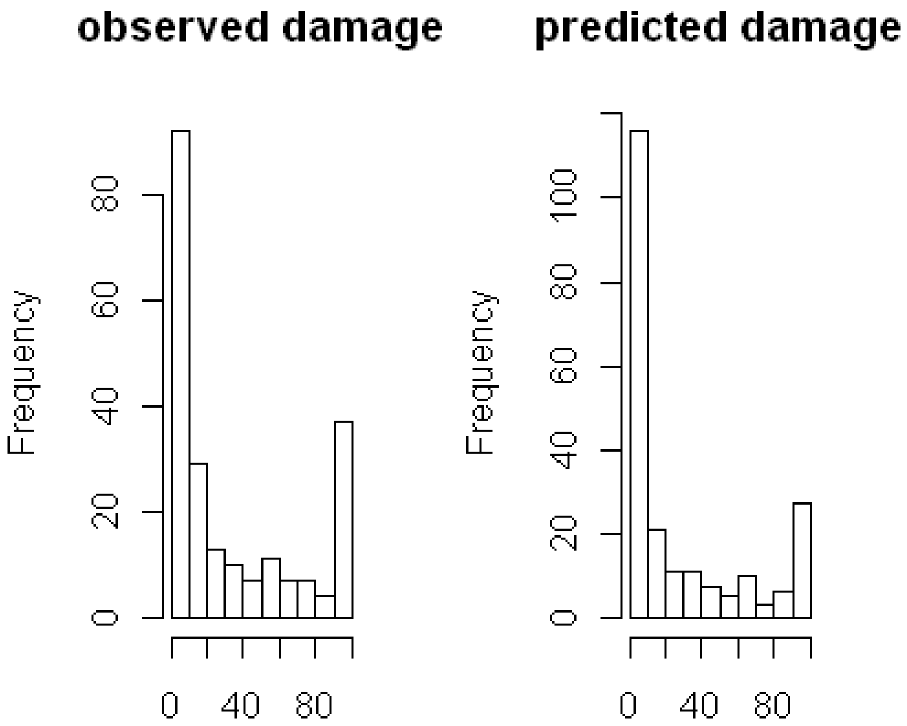
Résultats

	eff	MSE	biais²	SDSD	LCS
TR1	0.62	478	26	3	449
TR2	0.59	519	73	32	414
TR3	0.62	477	17	1	459
TR4	0.59	513	66	29	418

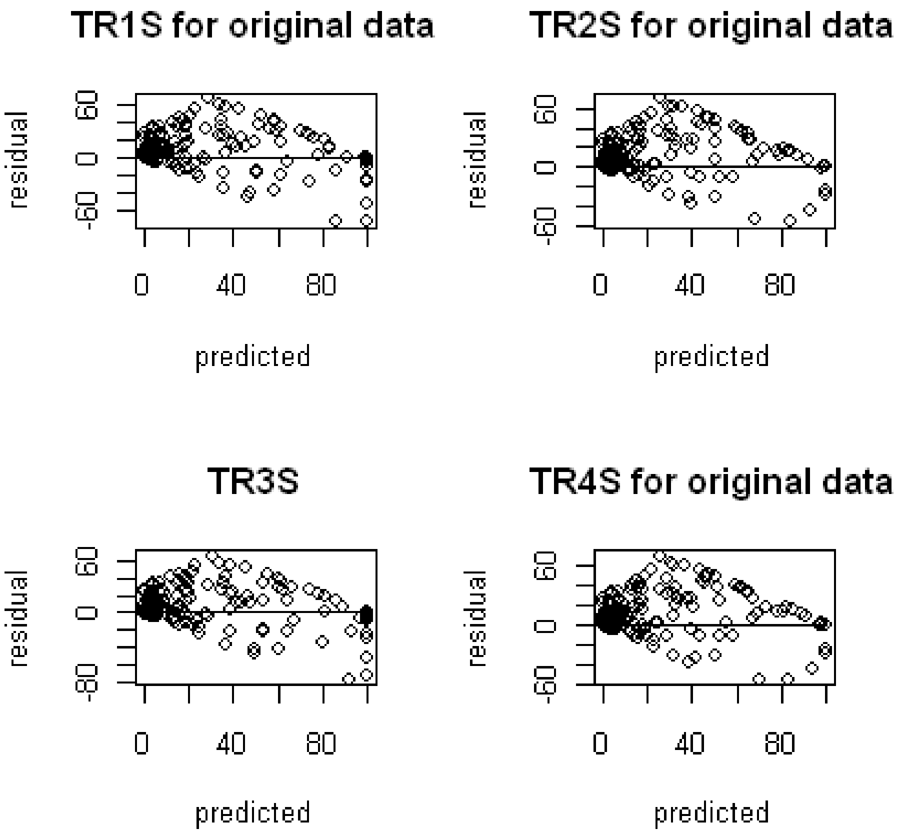
Erreur de prédiction

- En général, si on utilise les mêmes données pour estimer les paramètres et pour estimer MSE, alors MSE sous estime MSEP
- Mais l'erreur dépendra de p/n = 3/2001.
- On devrait pouvoir estimer MSEP par MSE

Données 2008



Erreurs de prédiction



Qualité de prédiction

	asin	asinDiff	dégâts	dégâtsDiff
efficience	0.73	0.68	0.73	0.69
MSEP	341	392	331	386
biais²	31	63	24	60
SDSD	9	37	5	36
LCS	302	293	303	291
nombre valeurs extrêmes	14	15	13	14

Intervalles de confiance

Calcul des intervalles de confiance

- La variance d'une prédiction

$$\hat{\text{var}}(\hat{y}) = z_i^T (\text{var}\,\hat{\theta}_{MV})z_i + \hat{\sigma}^2_{MV}$$
$$= \hat{\sigma}^2_{MV} z_i^T (Z^T Z)^{-1} z_i + \hat{\sigma}^2_{MV}$$

- L'intervalle de confiance

.

$$CI_\alpha = \left[\hat{y} - t_{n-p}^{\alpha/2} \sqrt{\hat{\text{var}}(\hat{y})}, \hat{y} + t_{n-p}^{\alpha/2} \sqrt{\hat{\text{var}}(\hat{y})} \right]$$

Hypothèses

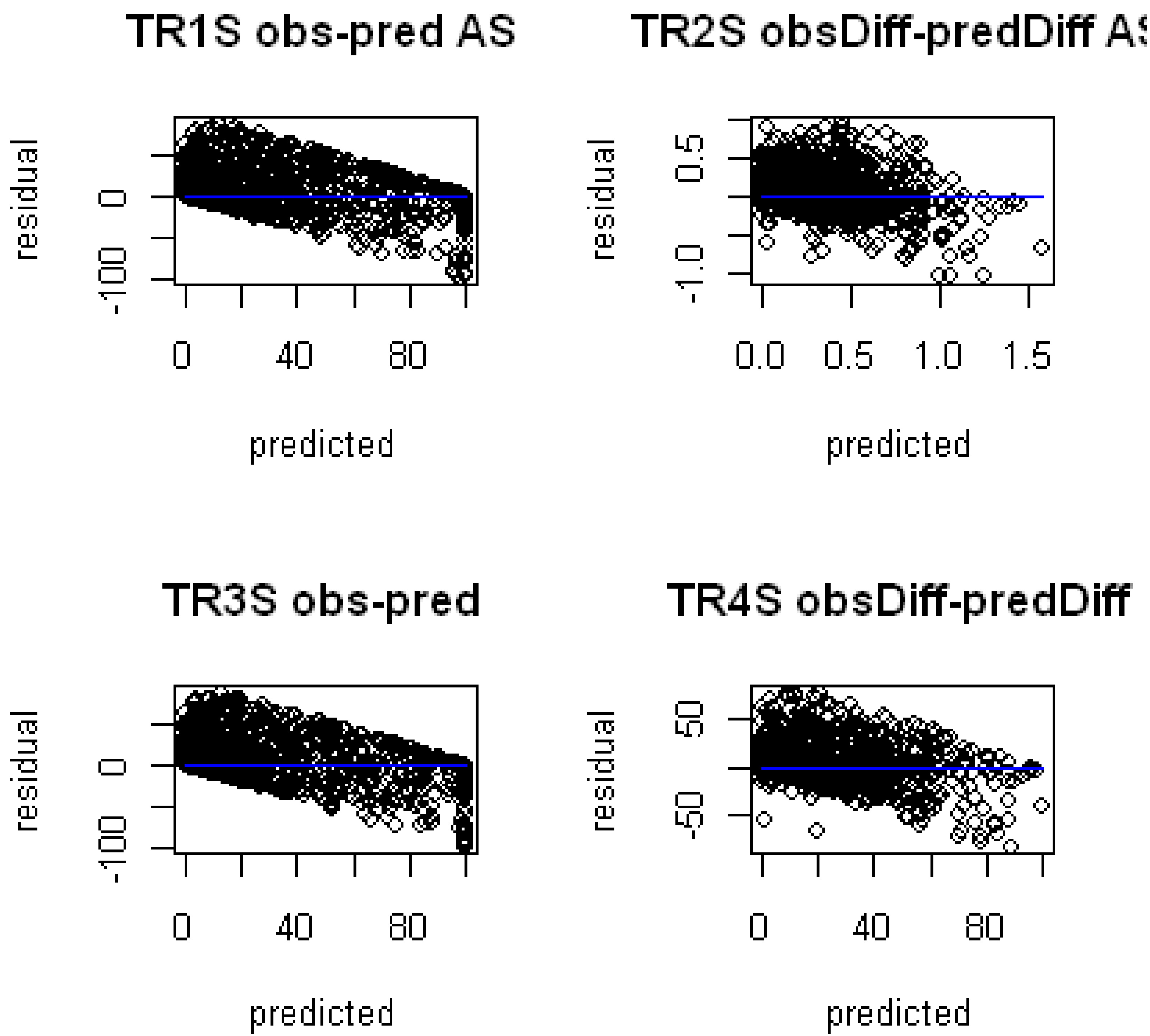
$$y_{sfi} = f_{fi}(X_{sfi};\theta) + \varepsilon_{sfi}$$

$E(\varepsilon_{sfi} \mid X_{sfi}) = 0$ espérance des erreurs=0

$\mathrm{var}(\varepsilon_{sfi} \mid X_{sfi}) = \sigma^2$ variance constante

$\varepsilon_{sfi} \mid X_{sfi} \sim N(0,\sigma^2)$ distribution normale

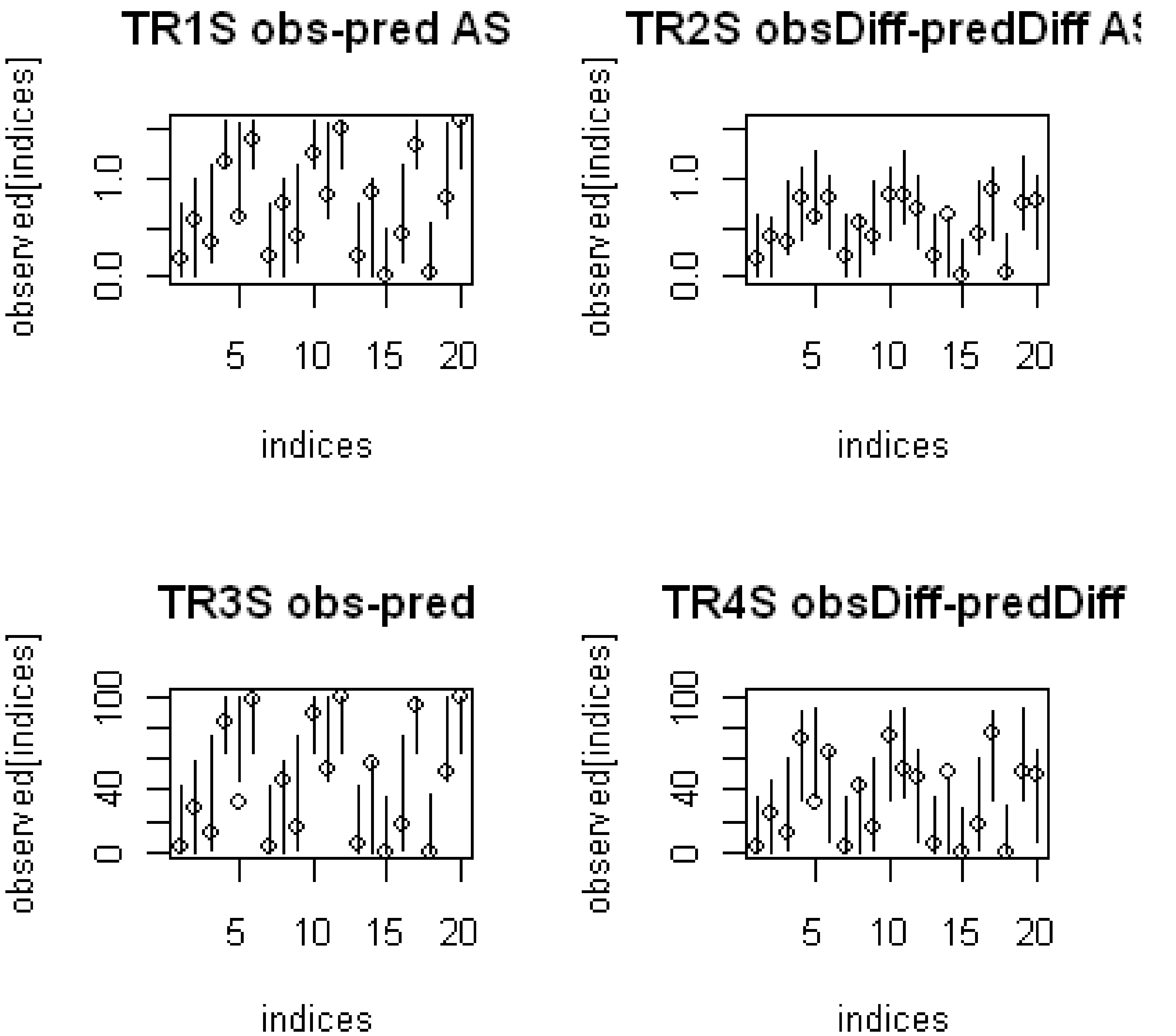
ε_{sfi} indépendants erreurs indépendantes



TR1S obs-pred AS
TR2S obsDiff-predDiff AS
TR3S obs-pred
TR4S obsDiff-predDiff
residual
predicted
-100
0
40
80
0.0
0.5
1.0
1.5
-1.0
-50
50

Conclusions sur hypothèses

- Toutes fausses


TR1S obs-pred AS
observed[indices]
0.0
1.0
5
10
15
20
indices
TR2S obsDiff-predDiff AS
observed[indices]
0.0
1.0
5
10
15
20
indices
TR3S obs-pred
observed[indices]
0
40
100
5
10
15
20
indices
TR4S obsDiff-predDiff
observed[indices]
0
40
100
5
10
15
20
indices

Fraction dans l'intervalle

	Fraction in 90% CI	ave obs value	ave length CI	min	max
TR1	0.86	0.51	69	0	1.57
TR2	0.87	0.26	53	0	1.57
TR3	0.89	30	45	0	100
TR4	0.86	17	38	0	100

Conclusions

- On traite le modèle comme un modèle de régression non-linéaire classique, avec 3 paramètres.

- Et le fait que c'est un modèle mécaniste?
  - Equations dynamiques.
    - Mais on regarde équations après intégration
  - Beaucoup de paramètres
    - Mais on n'a estimé que 3.
  - Beaucoup de variables explicatives
    - Mais on ne regarde que les prédictions (et non pas l'effet de chaque variable explicative)

THE END (for now)