Analyse de séries temporelles avec des modèles dynamiques linéaires en Agriculture.

Jeudi 10 juillet 2014 - Paris

Lucie MICHEL / lucie.michel@acta.asso.fr

Régression linéaire et non-linéaire

(linéaire, quadratique, cubique et linéaire + plateau)

Cours

Régression linéaire

Qu’est-ce qu’un modèle statistique ?

- **Un type de modèle mathématique particulier**

- **Un modèle qui inclut des éléments observables (les variables mesurées)**

… et des éléments non observables (les paramètres et certaines variables cachées)

- **Certains de ces éléments sont des variables aléatoires définies par des lois de probabilité.**

Qu'est-ce qu'un modèle statistique ?

Paramètre(s)

$$Y = f\left(X, \theta, \varepsilon\right)$$

Variable de réponse

Variable(s) explicative(s)

Résidu

Qu'est qu'un modèle linéaire ?

Paramètre(s)

$$Y = X\theta + \varepsilon$$

Variable de réponse

Variable(s) explicative(s)

Résidu

Qu'est-ce qu'un modèle linéaire ?

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1P} \\ x_{21} & x_{22} & \dots & x_{2P} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{NP} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \dots \\ \theta_P \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_N \end{pmatrix}$$

Vecteur des *N* observations de *Y*

Matrice des *N*P* valeurs des *P* variables explicatives

Vecteur des *P* paramètres

Vecteur des *N* termes résiduels

$$y_2 = x_{21}\theta_1 + x_{22}\theta_2 + \ldots + x_{2P}\theta_P + \varepsilon_2$$

Plusieurs types de modèles linéaires

- **Une variable explicative continue → Régression linéaire simple**
- **Plusieurs variables explicatives continues → Régression linéaire multiple**

- **Une variable explicative catégorielle → Analyse de variance à un facteur (ANOVA)**
- **Plusieurs variables explicatives catégorielles → Analyse de variance à 2, 3… facteurs**

- **Variables explicatives continues et catégorielles → Analyse de covariance (ANCOVA)**

A quoi peuvent servir ces modèles ?

- **Tester l'existence d'une relation entre la variable *Y* et une ou plusieurs variables explicatives (*X*)**

  → Test statistique

- **Quantifier l'effet de X sur Y**

  → Estimation et intervalle de confiance

- **Prédire *Y* en fonction de *X***

  → Prédiction

Autres types de modèles statistiques

- **Modèles linéaires généralisés (variable de réponse *Y* catégorielle)**

- **Modèles non-linéaires (*f* n'est pas linéaire)**

- **Modèles mixtes (données répétées non indépendantes)**

Pourquoi est-il souvent utilisé ?

- **Permet de modéliser beaucoup de phénomènes de manière réaliste**

- **Ses paramètres sont faciles à estimer**

- **Beaucoup d'outils statistiques sont associés aux modèles linéaires (tests)**

- **Attention : ses hypothèses ne sont pas toujours vérifiées.**

Etapes du développement d'un modèle

- **Définition des variables**
- **Définition des équations**
- **Estimation**
- **Tests et évaluation**
- **Utilisation**

Etapes du développement d'un modèle

- **Définition des variables**
- **Définition des équations**
- **Estimation**
- **Tests et évaluation**
- **Utilisation**

Exemple

- **BDD de rendement du colza (1950-2011)**

	A	B	C	D	E	F
1	Dpt_Num	Dpt_Nom	Year	Yield	Surface	Production
2	1	AIN	1950	6	600	3400
3	1	AIN	1951	10	500	5100
4	1	AIN	1952	8	600	4800
5	1	AIN	1953	8	380	3040
6	1	AIN	1954	10	450	4500
7	1	AIN	1955	12	400	4800
8	1	AIN	1956	9	2200	20000
9	1	AIN	1957	9	400	3600
10	1	AIN	1958	12	280	3362
11	1	AIN	1959	11	480	5160
12	1	AIN	1960	11	250	2700
13	1	AIN	1961	14	230	3120
14	1	AIN	1962	22	840	18400
15	1	AIN	1963	21	840	17520

Exemple

- **Evolution du rendement de colza dans l'Oise (60):**

- **Données:**

Dpt_Num	Dpt_Nom	Year	Yield	Surface	Production
60	OISE	1950	10	13000	130000
60	OISE	1951	15.00714286	14000	210100
60	OISE	1952	17	11000	187000
60	OISE	1953	13	4000	52000
60	OISE	1954	12	2200	26400
60	OISE	1955	17	1800	30600
60	OISE	1956	15	1000	15000
60	OISE	1957	17	2100	35700
60	OISE	1958	13	3000	39000
60	OISE	1959	18	1500	27000
60	OISE	1960	20	350	7000
60	OISE	1961	20	500	10000
60	OISE	1962	21.02803738	856	18000
60	OISE	1963	25	1830	45750
60	OISE	1964	24.52178649	1836	45022
60	OISE	1965	25.0797546	652	16352
60	OISE	1966	25	1300	32500
60	OISE	1967	27.73823884	2487	68985
60	OISE	1968	25.18382353	1632	41100
60	OISE	1969	23	1547	35581
60	OISE	1971	19.41165587	1081	20984
60	OISE	1972	19.35483871	434	8400
60	OISE	1973	25.36501901	263	6671
60	OISE	1974	30.8974359	195	6025
60	OISE	1975	22.760181	221	5030
60	OISE	1976	21.90163934	244	5344
60	OISE	1977	25.31372549	204	5164
60	OISE	1978	25.40037951	527	13386
60	OISE	1979	23.53415511	1947	45821
60	OISE	1980	29.93093728	2838	84944
60	OISE	1981	23.44615385	5200	121920
60	OISE	1982	26	5000	130000
60	OISE	1983	20	5970	119400
60	OISE	1984	30	3620	108600
60	OISE	1985	36	2945	106020
60	OISE	1986	32	5315	170080
60	OISE	1987	44	11340	498960
60	OISE	1988	35.78255675	16740	599000
60	OISE	1989	31.44	13100	411864

Exemple

- **Variable à expliquer:**
  - **Y: le rendement du colza dans l'Oise (4838 observations)**

- **Variables explicatives (selon le modèle utilisé):**
  - **t: le temps (l'année de récolte)**
  - **$t+t^2$: le temps + le temps au carré**
  - **$t+t^2+t^3$: le temps+ le temps au carré + le temps au cube.**

Etapes du développement d'un modèle

- **Définition des variables**
- **Définition des équations**
- **Estimation**
- **Tests et évaluation**
- **Utilisation**

Modèle linéaire

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_{4838} \end{pmatrix} = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \dots & \dots \\ 1 & t_{4838} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{4838} \end{pmatrix}$$

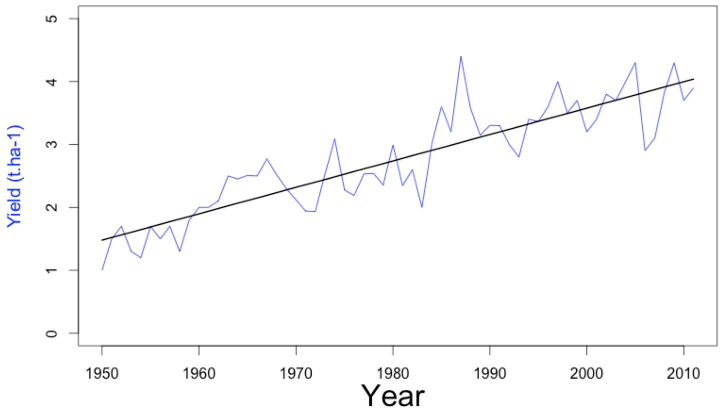
Vecteur des *4838* observations de Y

Matrice des *4838*2* valeurs (l'intercept et *1* variable explicative)

Vecteur des 2 paramètres

Vecteur des *4838* termes résiduels

Modèle linéaire



$$Y = \theta_1 + \theta_2 \times t + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

- **Dans R:**
  - **Fonction lm() ou glm():**

```
lineaire <- glm(Y~t)
```

Modèle quadratique

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_{4838} \end{pmatrix} = \begin{pmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \cdots & \cdots & \cdots \\ 1 & t_{4838} & t_{4838}^2 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_{4838} \end{pmatrix}$$

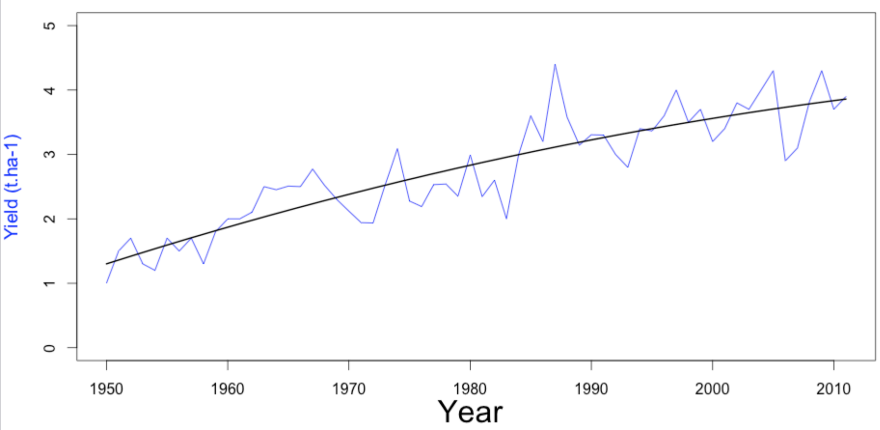
Vecteur des *4838* observations de Y

Matrice des *4838* 3* valeurs (intercept et 2 variables explicatives)

Vecteur des *3* paramètres

Vecteur des *4838* termes résiduels

Modèle quadratique



$$Y = \theta_1 + \theta_2 \times t + \theta_3 \times t^2 + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

- **Dans R:**
  - **Fonction lm() ou glm():**

lineaire <- glm(Y~t+t2)

avec t2: le temps au carré

Modèle cubique

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_{4838} \end{pmatrix} = \begin{pmatrix} 1 & t_1 & t_1^2 & t_1^3 \\ 1 & t_2 & t_2^2 & t_2^3 \\ \dots & \dots & \dots & \dots \\ 1 & t_{4838} & t_{4838}^2 & t_{4838}^3 \end{pmatrix} + \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{4838} \end{pmatrix}$$

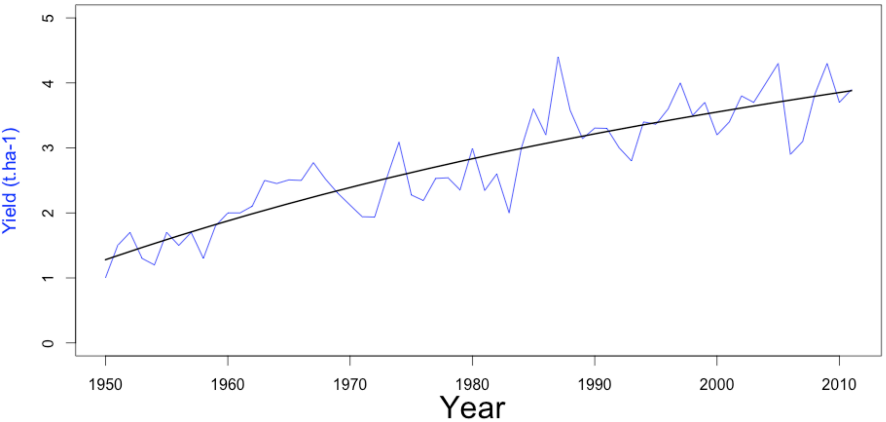
Vecteur des *4838* observations de Y

Matrice des *4838* 4* valeurs (intercept et *3* variables explicatives)

Vecteur des *4838* termes résiduels

Vecteur des *4* paramètres

Modèle cubique



$$Y = \theta_1 + \theta_2 \times X + \theta_3 \times t^2 + \theta_4 \times t^3 + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

- **Dans R:**
  - **Fonction lm() ou glm():**

lineaire <- glm(Y~t+t2+t3)

avec t2: le temps au carré

avec t3: le temps au cube

Etapes du développement d'un modèle

- **Définition des variables**
- **Définition des équations**
- **Estimation**
- **Tests et évaluation**
- **Utilisation**

Estimation

Objectif : trouver des valeurs des paramètres qui ont de bonnes propriétés

Biais : écart entre la vrai valeur d'un paramètre et l'espérance ('moyenne') de l'estimation

Variance : variance de l'estimation sur l'ensemble des jeux de données possibles

Méthodes des moindres carrés ordinaires

Trouver les valeurs des paramètres minimisant

$$SCR = \|Y - X\theta\|^2$$

La solution est

$$\theta_{OLS} = \left(X'X\right)^{-1} X'Y$$

Méthodes des moindres carrés ordinaires

Dans le cas de la régression linéaire simple, on minimise

$$SCR = \sum_{i=1}^{N}\left[y_i - \left(\alpha + \beta x_i\right)\right]^2$$

Propriétés

L'estimateur OLS est, sous certaines conditions, sans biais (biais=0) et de variance minimale

$$E\left(\theta_{OLS}\right) = \theta$$

$$\operatorname{var}\left(\theta_{OLS}\right) = \left(X'X\right)^{-1}\operatorname{var}\left(\varepsilon\right)$$

Test d’égalité à zéro d’un paramètre

(test de student)

Test d'égalité à zéro d'un paramètre

H_0 : « $\theta_i = 0$ » contre H_1 : « $\theta_i \neq 0$ »

Utile pour déterminer si une variable explicative à un effet sur la variable de réponse

Test d'égalité à zéro d'un paramètre

H_0 : « $\theta_i = 0$ » contre H_1 : « $\theta_i \neq 0$ »

On va chercher à

limiter le risque de se tromper en rejetant H_0

c'est à dire

limiter le risque de conclure faussement à l'existence d'un effet (risque de 1er espèce)

Test d'égalité à zéro d'un paramètre

H_0 : « $\theta_i = 0$ » contre H_1 : « $\theta_i \neq 0$ »

On utilise une statistique de test T calculée à partir du modèle et des données.

On rejette H_0 si $T>T^*$.

Exemple (linéaire)

- **Evolution du rendement de colza dans l'Oise (60):**

- **Script:**

```
TAB_Yield <- read.table("COLZA_DPT_surface_prod_rdt_R.txt", header = T)
Dpt_Nom <- levels(TAB_Yield$Dpt_Nom)


Dpt_i <- "OISE"
TAB_Yield_i<-TAB_Yield[TAB_Yield$Dpt_Nom == Dpt_i,]

Year <- TAB_Yield_i$Year

Yield <- TAB_Yield_i$Yield/10
lineaire <- glm(Yield~Year)
print(summary(lineaire ))
```

Exemple (linéaire)

- **Evolution du rendement de colza dans l'Oise (60):**
- **Résultats:**

```
Call:
glm(formula = Yield ~ Year)
```

Distribution des résidus

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.92866  -0.25160   0.01177   0.21072   1.36875
```

Estimation des paramètres
Test de T
P-value

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -80.36157    5.70482  -14.09   <2e-16 ***
Year          0.04197    0.00288   14.57   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1637744)
```

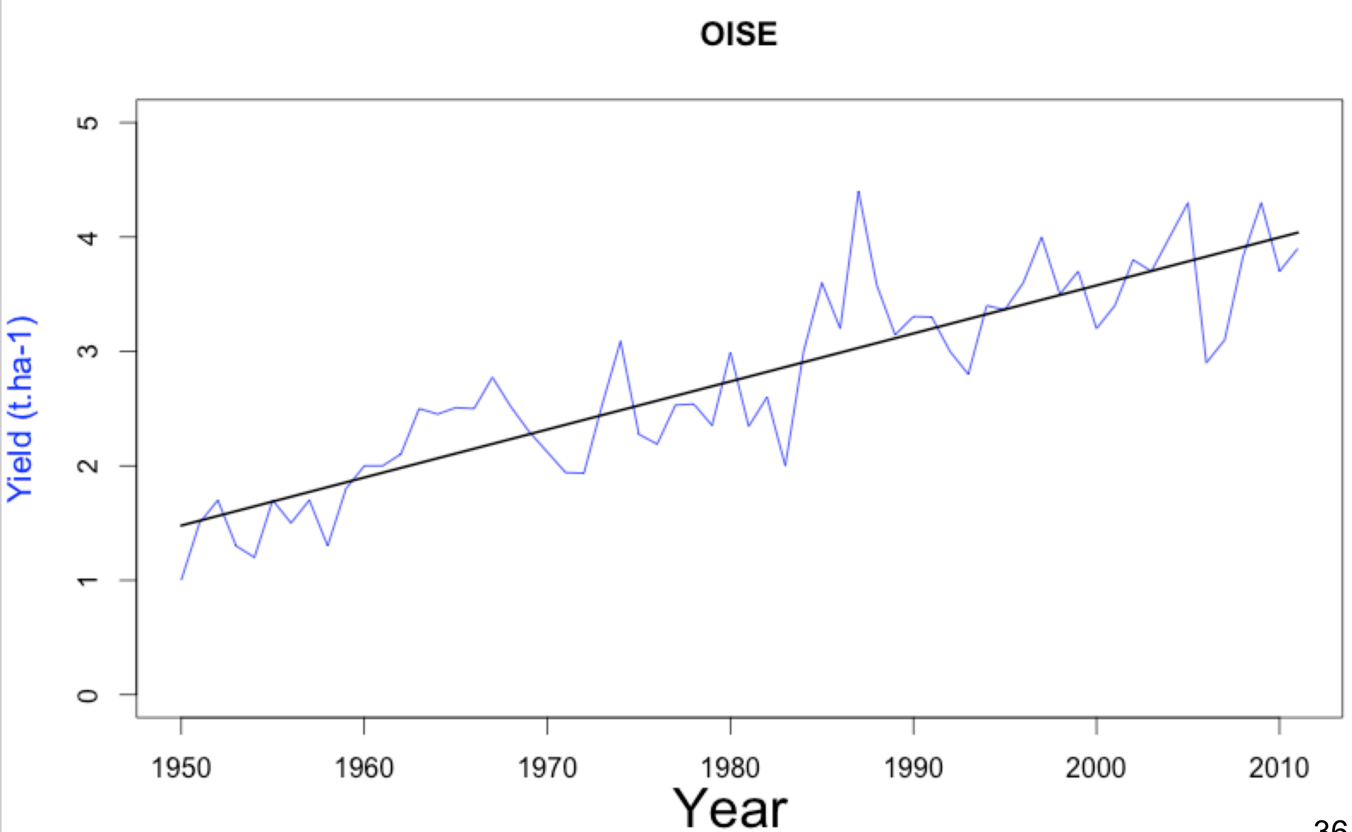
Variance
Variance des résidus
AIC

```
    Null deviance: 44.4391  on 60  degrees of freedom
Residual deviance:  9.6627  on 59  degrees of freedom
AIC: 66.712

Number of Fisher Scoring iterations: 2
```

Exemple (linéaire)

```
plot(Year, Yield, xlab="", ylab="Yield (t.ha-1)", type ="l", lwd=1, xlim=c(1950,2011),
     ylim=c(0,max(Yield)), cex.lab=1.2,col = "blue", col.lab = "blue")
lines(Year, predict(lineaire), lwd=2)
mtext("Year", side =1, line=2.5, at=1980, cex=2)
title(Dpt_i)
```

- **Graphique:**



Exemple (quadratique)

- **Evolution du rendement de colza dans l'Oise (60):**

- **Script:**

```
#Rejutement des annees pour les modeles cubique et quadratique
Year_b <- Year - Year[1]
#Annes pour le modele cubique
Year2_b<-Year_b*Year_b

#Model quadratique
quadratiq <- glm(Yield~Year_b+Year2_b)
print(summary(quadratiq))
```

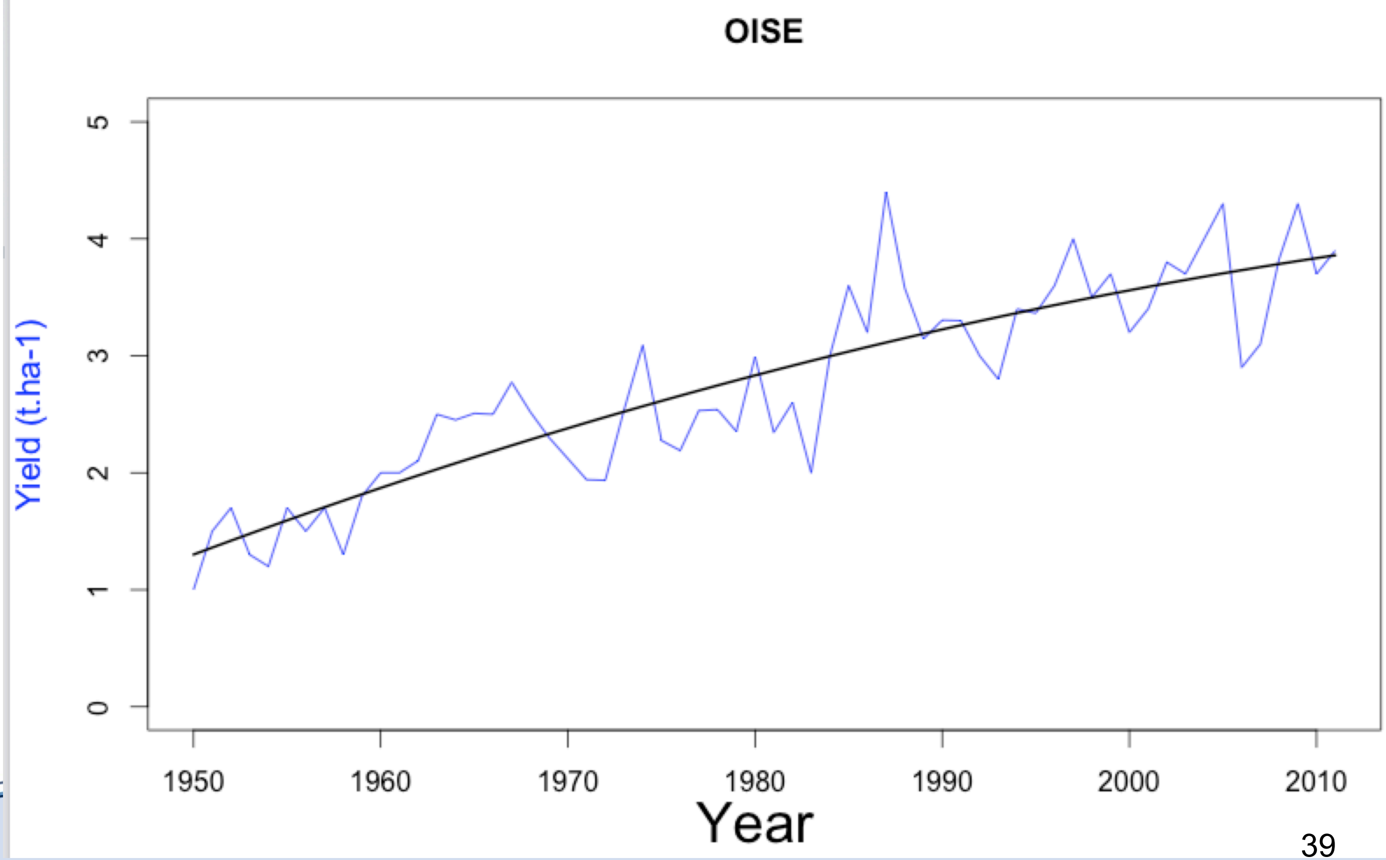
Exemple (quadratique)

- **Evolution du rendement de colza dans l'Oise (60):**
- **Résultats:**

```
Call:
glm(formula = Yield ~ Year_b + Year2_b)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.95652  -0.29694   0.03387   0.18062   1.28631

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.3012657  0.1472275   8.838 2.45e-12 ***
Year_b       0.0598512  0.0111865   5.350 1.56e-06 ***
Year2_b     -0.0002937  0.0001777  -1.653    0.104
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1591058)

    Null deviance: 44.4391  on 60  degrees of freedom
Residual deviance:  9.2281  on 58  degrees of freedom
AIC: 65.905

Number of Fisher Scoring iterations: 2
```

Exemple (quadratique)

- **Graphique:**



Exemple (cubique)

- **Evolution du rendement de colza dans l'Oise (60):**

- **Script:**

```
#Annees pour le modele cubique
Year3_b <- Year2_b*Year_b

#Model cubique
cubiq <- glm(Yield~Year_b+Year2_b+Year3_b)
print(summary(cubiq))
```

Exemple (cubique)

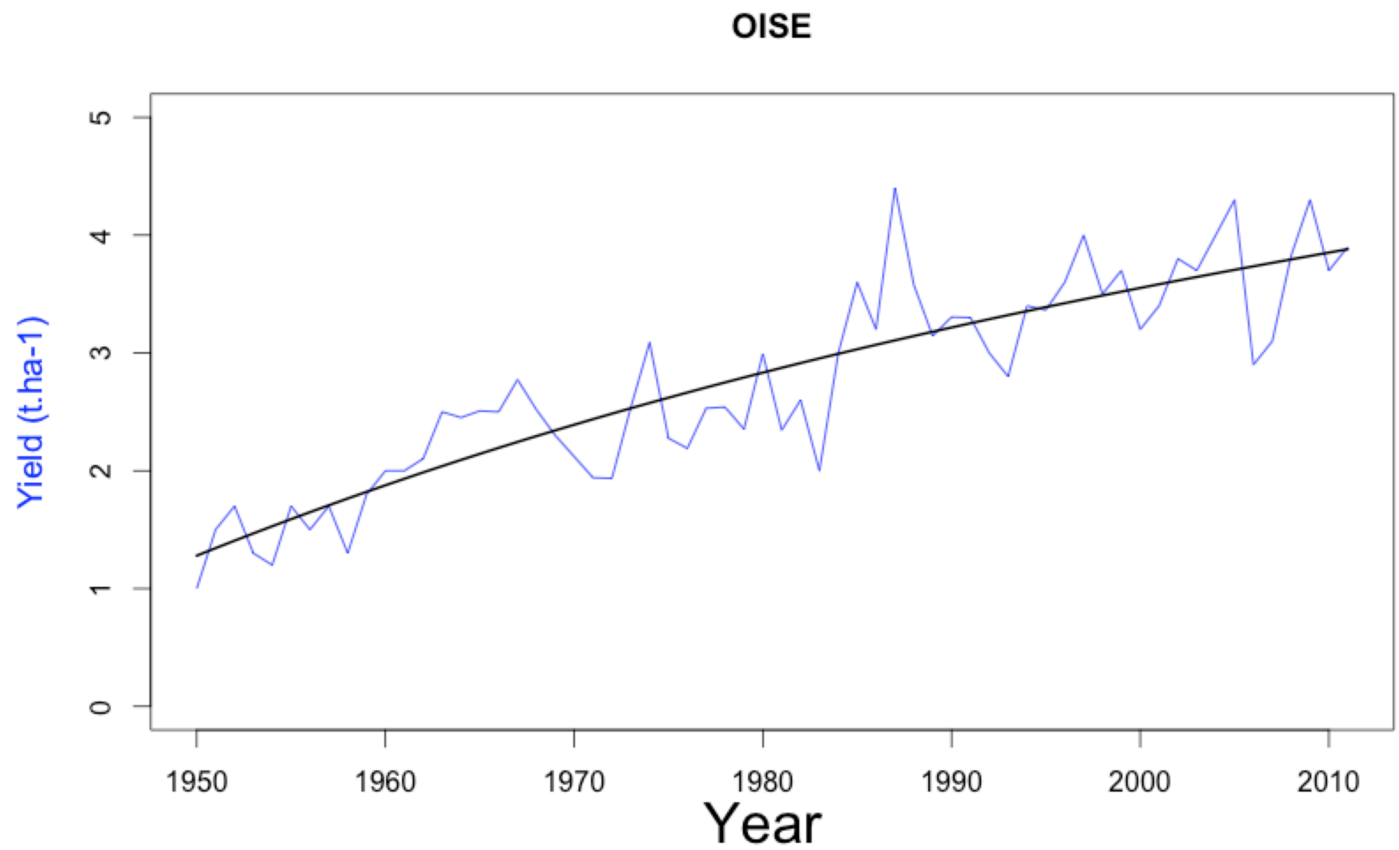
- **Evolution du rendement de colza dans l'Oise (60):**

- **Résultats:**

```
Call:
glm(formula = Yield ~ Year_b + Year2_b + Year3_b)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9539  -0.2797   0.0192   0.1855   1.2930

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.280e+00  1.931e-01   6.626 1.33e-08 ***
Year_b       6.432e-02  2.791e-02   2.304   0.0249 *
Year2_b     -4.778e-04  1.068e-03  -0.447   0.6562
Year3_b      2.007e-06  1.148e-05   0.175   0.8618
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1618103)

    Null deviance: 44.4391  on 60  degrees of freedom
Residual deviance:  9.2232  on 57  degrees of freedom
AIC: 67.872

Number of Fisher Scoring iterations: 2
```

Exemple (cubique)

- **Graphique:**



Etapes du développement d'un modèle

- **Définition des variables**
- **Définition des équations**
- **Estimation**
- **Tests et évaluation**
- **Utilisation**

Evaluation

- **Tests**
- **Analyse de résidus**
- **Critères d'évaluation**

Le coefficient de détermination R²

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - y_i^{pred}\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2}$$

$$R^2_{ajust} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - y_i^{pred}\right)^2 / \left(N - P - 1\right)}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2 / \left(N - 1\right)}$$

Critères d'évaluation

➢ Le critère d'Akaïke

$$AIC = -2\ln\left(L\right) - 2\left(P+1\right)$$

Hirotugu Akaike (né en 1927 au Japon)

Hirotugu Akaike was born in 1927 in Fujinomiya-shi, Shizuoka-jen, in Japan. He received B.S. and D.S. degrees in mathematics from the University of Tokyo in 1952 and 1961, respectively. He worked at the Institute of Statistical Mathematics for over 30 years, becoming its Director General in 1982. He has received many awards, prizes, and honors for his work in theoretical and applied statistics (deLeeuw 1992, Parzen 1994). The three-volume set, "*Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach* (Bozdogan 1994) commemorated Professor Hirotugu Akaike's 65th birthday. Bozdogan (1994) records that the idea of a connection between the Kullback–Leibler discrepancy and the empirical log-likelihood function occurred to Akaike on the morning of March 16, 1971, as he was taking a seat on a commuter train.

4.2.2 Averaging Across Model Parameters

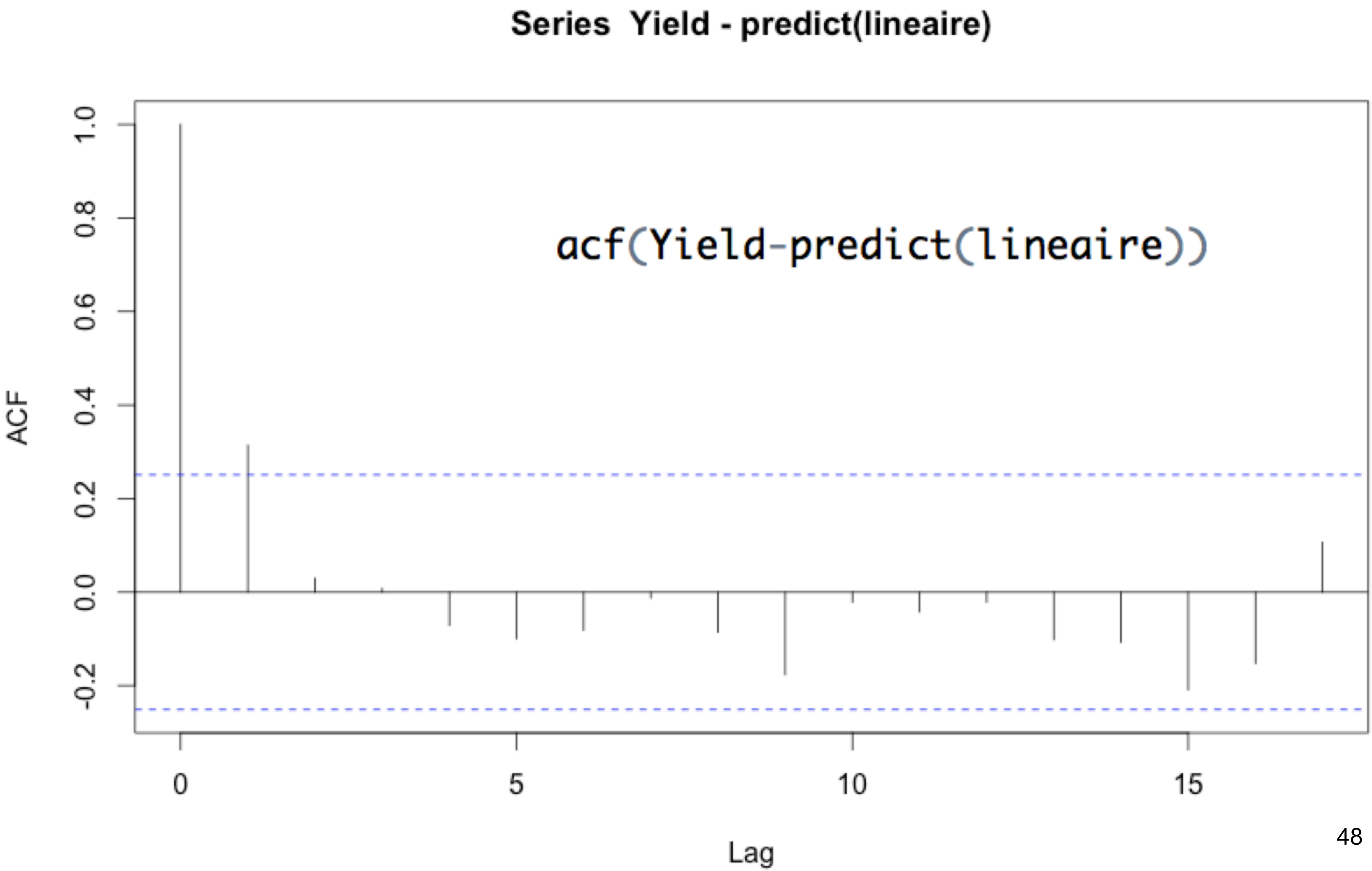
If one has a large number of closely related models, such as in linear-regression based variable selection (e.g., all subsets selection), designation of a single best model is unsatisfactory because that "best" model is often highly variable. That is, the model estimated to be best would vary from data set to data set, where replicate data sets would be collected under the same underlying process. In this situation, model averaging provides a relatively much more stabilized inference.

The concept of inference being tied to all the models can be used to reduce model selection bias effects on linear regression coefficient estimates in all subsets selection. For the linear regression coefficient β_j associated with predictor variable x_j there are two versions of model averaging. First, we have the estimate $\hat{\bar{\beta}}_j$ where β_j is averaged over all models in which x_j appears (i.e.,

Test: Exemple (linéaire)

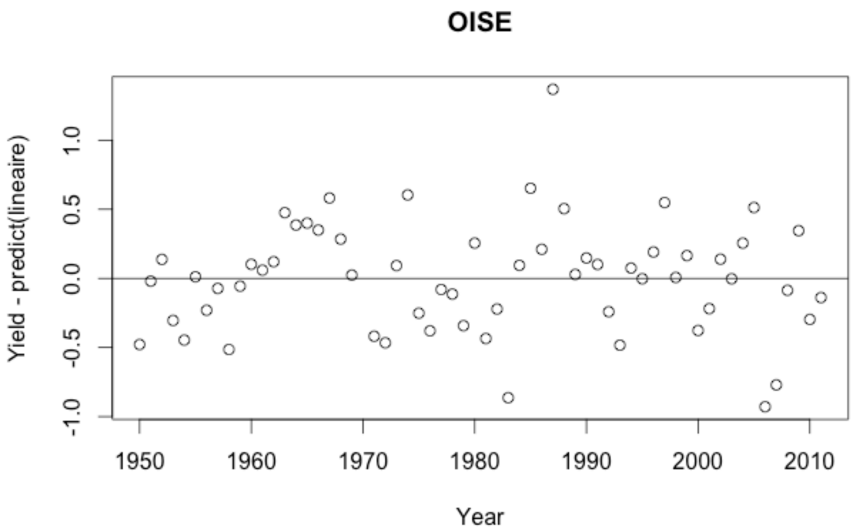
```
Call:
glm(formula = Yield ~ Year)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.92866  -0.25160   0.01177   0.21072   1.36875

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -80.36157    5.70482  -14.09   <2e-16 ***
Year          0.04197    0.00288   14.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1637744)

    Null deviance: 44.4391  on 60  degrees of freedom
Residual deviance:  9.6627  on 59  degrees of freedom
AIC: 66.712

Number of Fisher Scoring iterations: 2
```

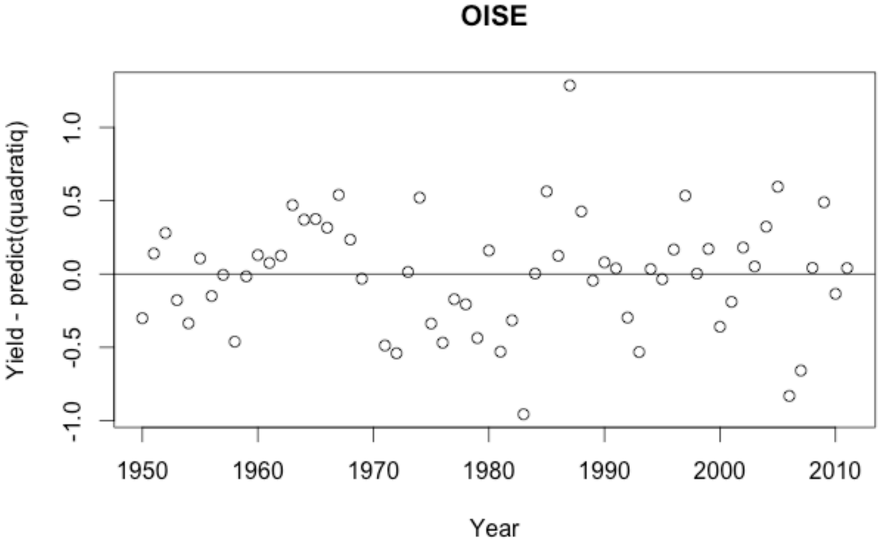
Autocorrélation : Exemple (linéaire)

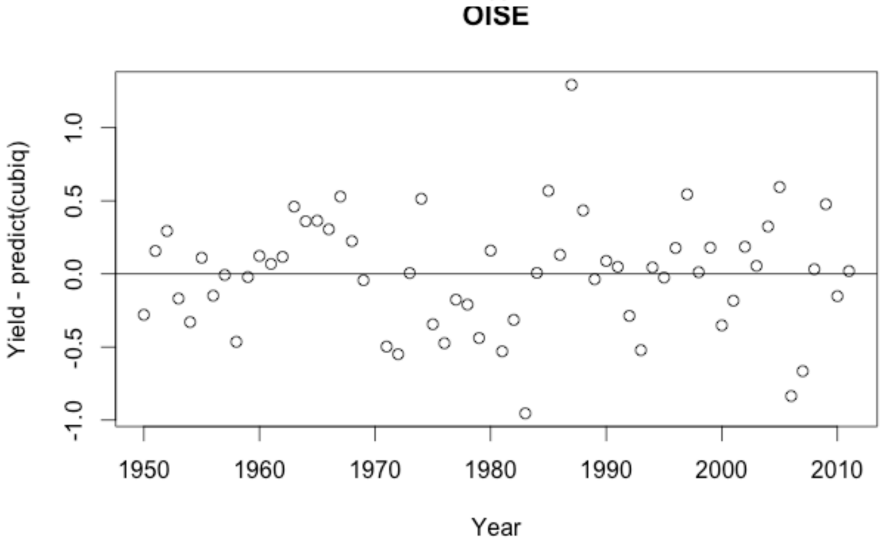


Graphique des résidus: Exemple

```
#Graphique des résidus pour le modele lineaire
plot(Year,Yield-predict(lineaire))
abline(0,0)
title(Dpt_i)
```







Critère d'évaluation R^2: Exemple

	Linéaire	Quadratique	Cubique
R^2	0.7826	0.7923	**0.7925**

```
R_lineaire <-1-((sum((Yield-predict(lineaire))^2))/(sum((Yield-mean(Yield))^2)))
```

```
Call:
glm(formula = Yield ~ Year)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.92866  -0.25160   0.01177   0.21072   1.36875

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -80.36157    5.70482  -14.09   <2e-16 ***
Year          0.04197    0.00288   14.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1637744)

    Null deviance: 44.4391  on 60  degrees of freedom
Residual deviance:  9.6627  on 59  degrees of freedom
AIC: 66.712

Number of Fisher Scoring iterations: 2
```

$$R^2 = 1 - \frac{\text{Residual deviance}}{\text{Null deviance}}$$

Critère d'évaluation
AIC: Exemple

	Linéaire	Quadratique	Cubique
AIC	66,71	**65,91**	67,87

Régression non linéaire

Modèle non linéaire

Paramètre(s)

$$Y = f\left(X,\theta\right) + \varepsilon$$

Variable de réponse

Variable(s) explicative(s)

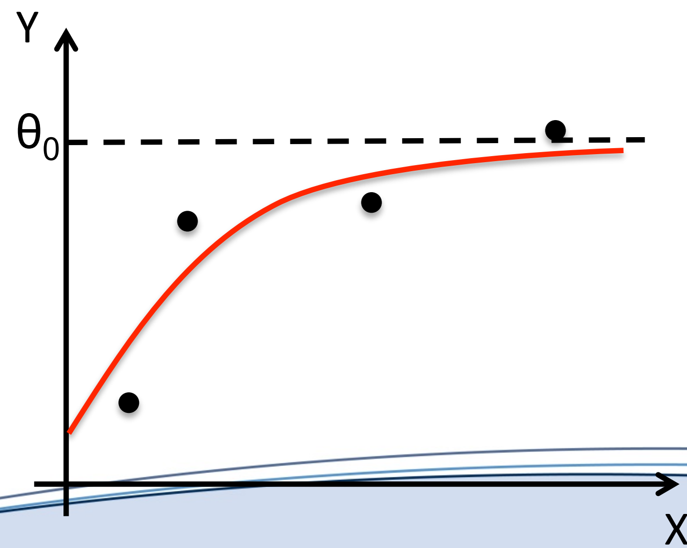
Résidu

Fonction non linéaire= combinaison de paramètres non linéaire

Modèle non linéaire

- **Fonction exponentielle:**

$$Y = \theta_0 \left(1 - \exp^{-\theta_1 \left(X + \theta_2 \right)} \right) + \varepsilon$$



Etapes du développement d'un modèle

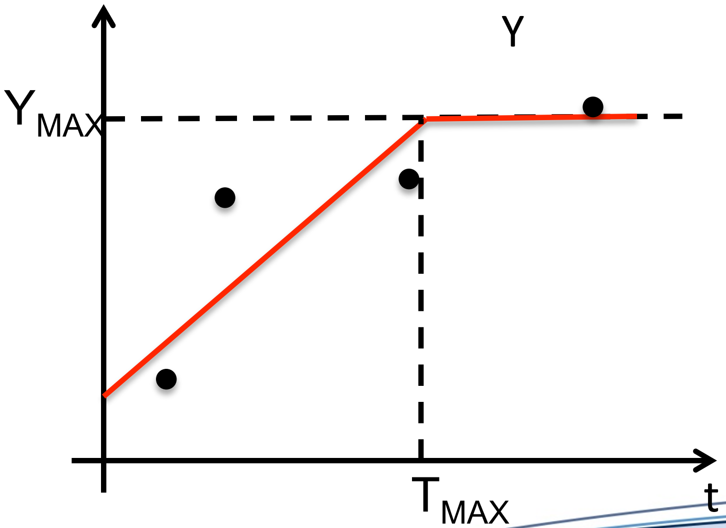
- **Définition des variables**
- **Définition des équations**
- **Estimation**
- **Tests et évaluation**
- **Utilisation**

Modèle linéaire + plateau: Equation

$$Y = Y_{\max} + P \times \left(t - T_{MAX}\right) + \varepsilon \quad \text{si } t<T_{MAX}$$

$$Y = Y_{MAX} + \varepsilon \quad \text{sinon}$$

$$\varepsilon \sim N(0,\sigma^2)$$



- **Dans R:**
  - **Fonction nls():**

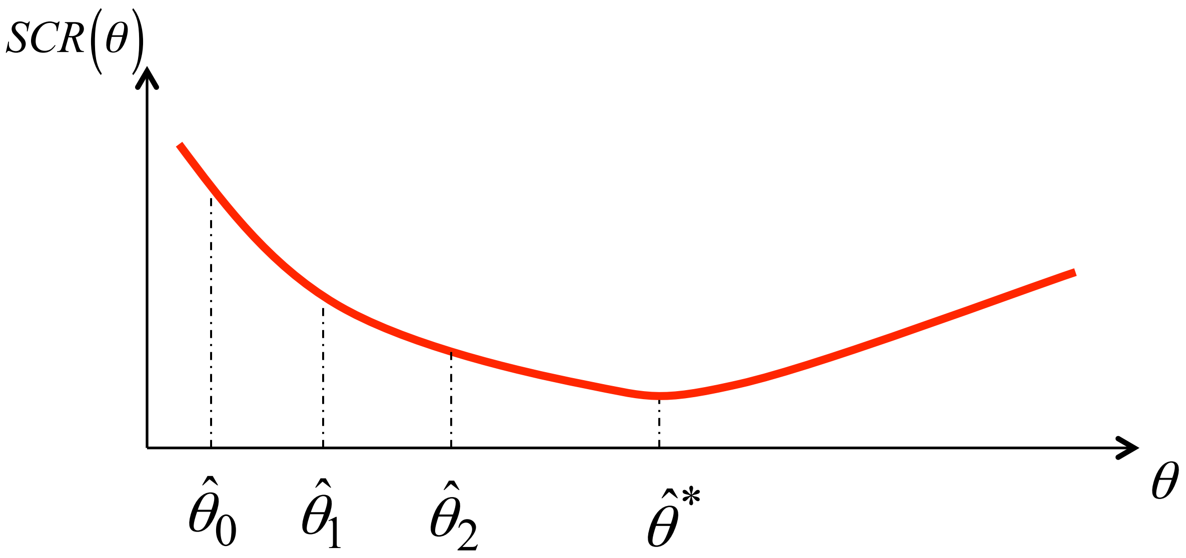
Etapes du développement d'un modèle

- **Définition des variables**
- **Définition des équations**
- **Estimation**
- **Tests et évaluation**
- **Utilisation**

Modèle non linéaire: estimation

– **Méthode des moindres carrés**

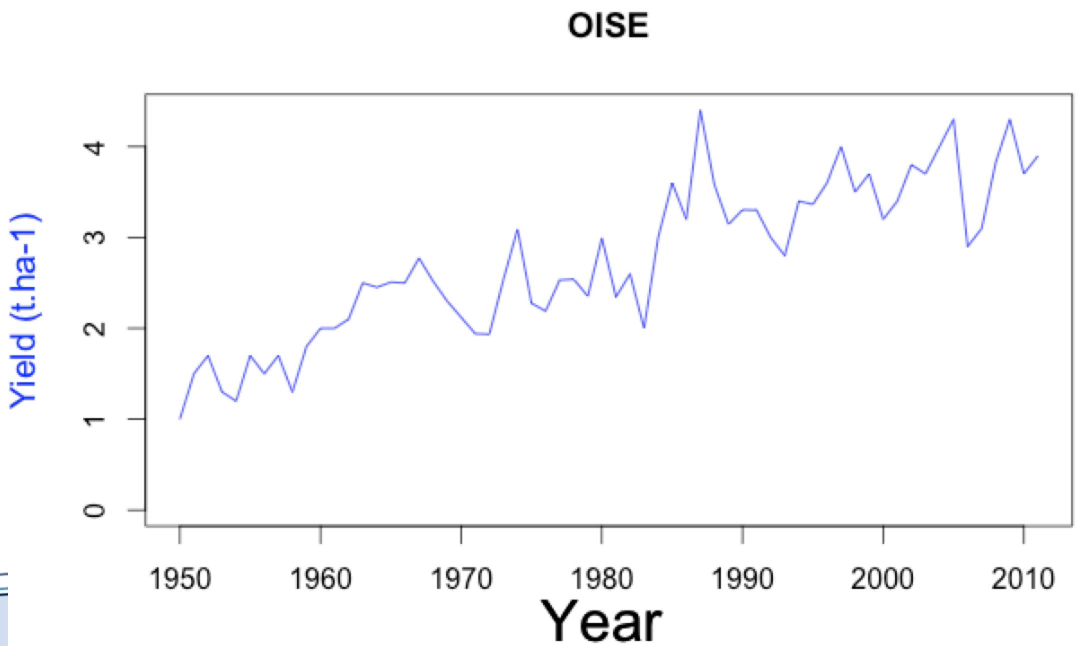
Local and global optimum



Modèle linéaire + plateau: paramètres

- Y_{MAX}: rendement maximal atteint lors du plateau
- T_{MAX}: année d'atteinte du plateau
- P: pente de la partie linéaire avant le plateau

- Exemple pour l'Oise:
  - Ymax =4
  - Tmax =1998
  - P=0.05



Exemple (linéaire +plateau)

- **Evolution du rendement de colza dans l'Oise (60):**

- **Script:**

```
#Fonction pour le modele lineaire+plateau
LP<- function (t, Ymax, Tmax, P) {
  Y <- Ymax+P*(t-Tmax)
  Y[Y>Ymax] <- Ymax
  Y
}

#Liste des 3 paramètres pour le modele voullu
list1<- list(Ymax= 4, P= 0.05, Tmax= 1998)
print(list1)

#Modele lineaire+plateau
  lplateau <- nls(Yield ~LP(Year, Ymax, Tmax, P),start= list1, trace= T)
  print(summary( lplateau))
```

Exemple (linéaire +plateau)

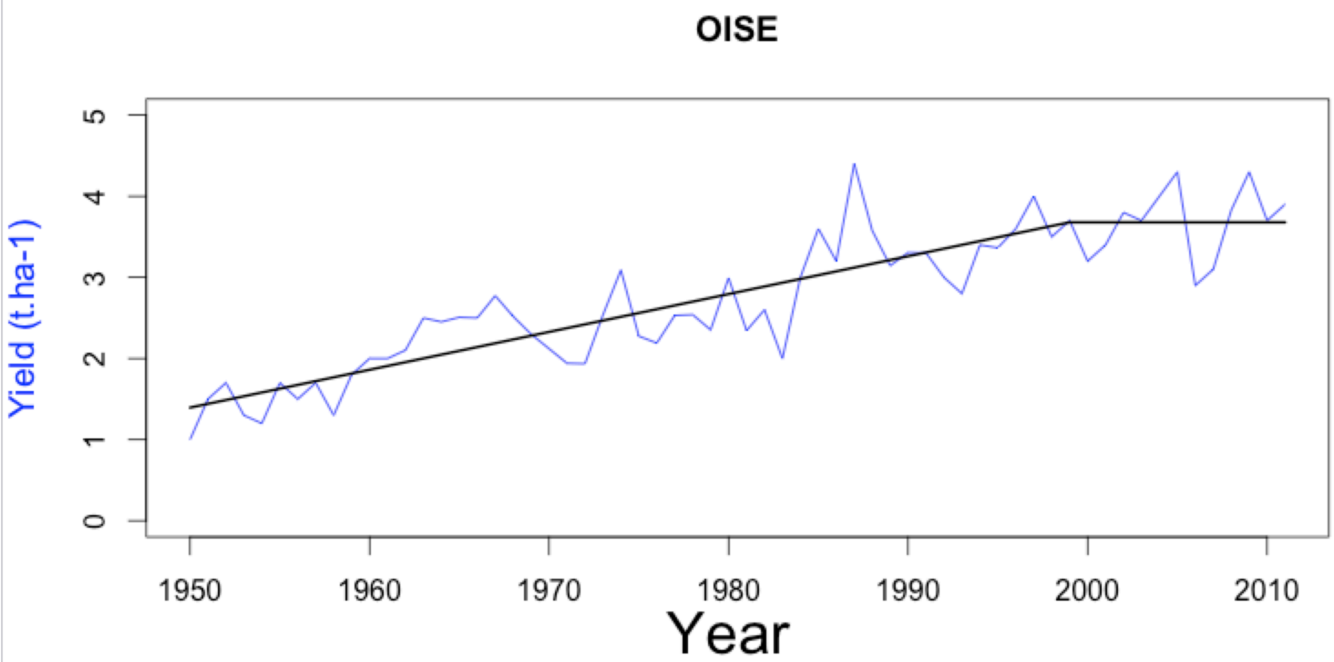
- **Evolution du rendement de colza dans l'Oise (60):**

- **Résultats:**

```
> #Modele lineaire+plateau
> lplateau <- nls(Yield ~LP(Year, Ymax, Tmax, P),start= list1, trace= T)
14.73323 :      4.00     0.05 1998.00
9.362422 :  3.678980e+00 4.665554e-02 1.998896e+03
9.36199 :  3.678980e+00 4.665554e-02 1.998961e+03
> print(summary( lplateau))

Formula: Yield ~ LP(Year, Ymax, Tmax, P)

Parameters:
      Estimate Std. Error t value Pr(>|t|)
Ymax 3.679e+00  1.114e-01   33.02   <2e-16 ***
P    4.666e-02  4.062e-03   11.49   <2e-16 ***
Tmax 1.999e+03  3.455e+00  578.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4018 on 58 degrees of freedom

Number of iterations to convergence: 2
Achieved convergence tolerance: 3.494e-09

> print(AIC(lplateau))
[1] 66.78333
```

Exemple (linéaire +plateau)

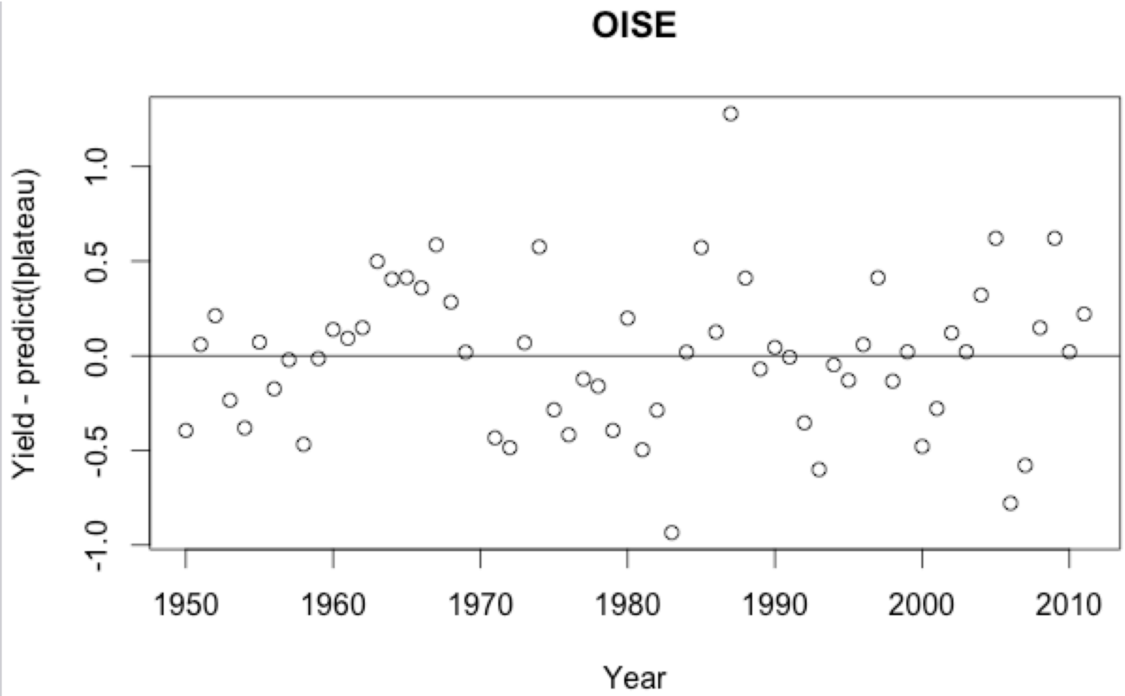
- **Graphique:**



Etapes du développement d'un modèle

- **Définition des variables**
- **Définition des équations**
- **Estimation**
- **Tests et évaluation**
- **Utilisation**

Evaluation (linéaire + plateau): Exemple



R^2	0.7893
AIC	66,78

Comparaison modèles

	Linéaire	Quadratique	Cubique	Linéaire + plateau
R^2	0.7825	0.7923	**0.7924**	0,7893
AIC	66,712	**65,905**	67,872	66,783

Régression linéaire et non-linéaire (linéaire, Q, C, LP)

TD

Enoncé

- **A partir de la BDD blé:**
  - **Pour les départements: AIN,EURE (et CREUSE).**
    - **Faire tourner les 4 modèles (linéaire, quadratique, cubique, linéaire + plateau)**

	Y_{MAX}	P	T_{MAX}
AIN	7	0.11	1998
EURE	8.5	0.12	1998
CREUSE	5.5	0.10	1996

    - **Récupérer les sorties des modèles (estimation, AIC)**
    - **Calculer les R^2**
    - **Faire le graphique des résidus**
    - **Faire le graphique de l'évolution du rendement**
    - **Comparer les modèles**

AIN

```
> print(summary(lineaire))

Call:
glm(formula = Yield ~ Year)

Deviance Residuals:
    Min        1Q    Median        3Q       Max
-1.3496   -0.3926    0.1013    0.3196    1.4630

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.987e+02  8.651e+00  -22.97   <2e-16 ***
Year         1.025e-01  4.368e-03   23.48   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.

(Dispersion parameter for gaussian family taken to be 0.

    Null deviance: 231.463  on 61  degrees of freedom
Residual deviance:  22.726  on 60  degrees of freedom
AIC: 119.72

Number of Fisher Scoring iterations: 2
```

```
> print(summary(quadratiq))

Call:
glm(formula = Yield ~ Year_b + Year2_b)

Deviance Residuals:
    Min        1Q    Median        3Q       Max
-1.3552   -0.3900    0.1106    0.3270    1.4634

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.234e+00  2.290e-01   5.387 1.31e-06 ***
Year_b      1.007e-01  1.736e-02   5.801 2.75e-07 ***
Year2_b     2.995e-05  2.753e-04   0.109    0.914
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0

(Dispersion parameter for gaussian family taken to be

    Null deviance: 231.463  on 61  degrees of freedom
Residual deviance:  22.722  on 59  degrees of freedom
AIC: 121.71

Number of Fisher Scoring iterations: 2
```

AIN

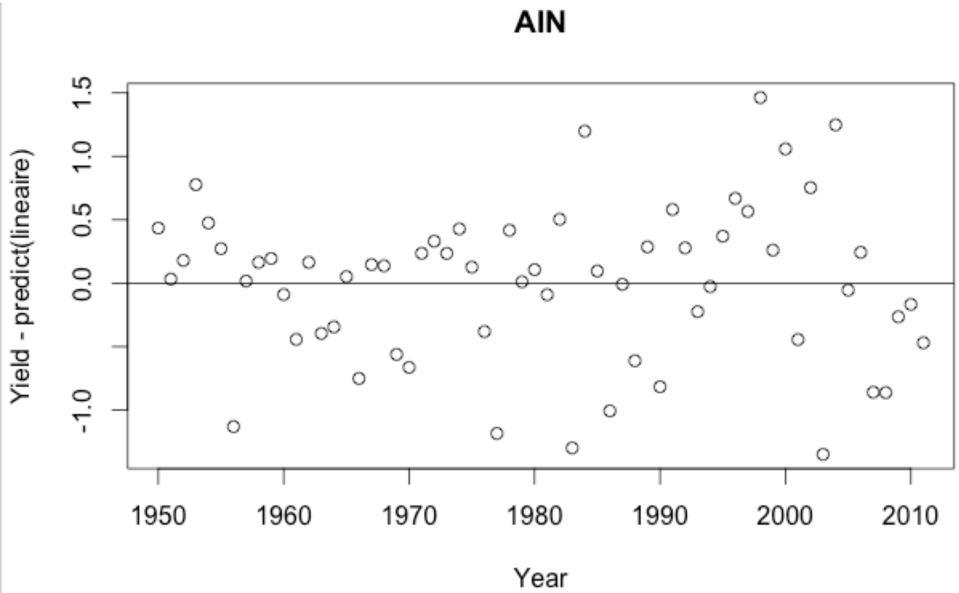
```
> print(summary(cubiq))

Call:
glm(formula = Yield ~ Year_b + Year2_b + Year3_b)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.42030  -0.28254   0.06452   0.35947   1.26814

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.680e+00  2.852e-01   5.891 2.06e-07 ***
Year_b       9.166e-03  4.082e-02   0.225   0.8231
Year2_b      3.812e-03  1.562e-03   2.440   0.0178 *
Year3_b     -4.134e-05  1.683e-05  -2.456   0.0170 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.

(Dispersion parameter for gaussian family taken to be 0

    Null deviance: 231.463  on 61  degrees of freedom
Residual deviance:  20.581  on 58  degrees of freedom
AIC: 117.58

Number of Fisher Scoring iterations: 2
```

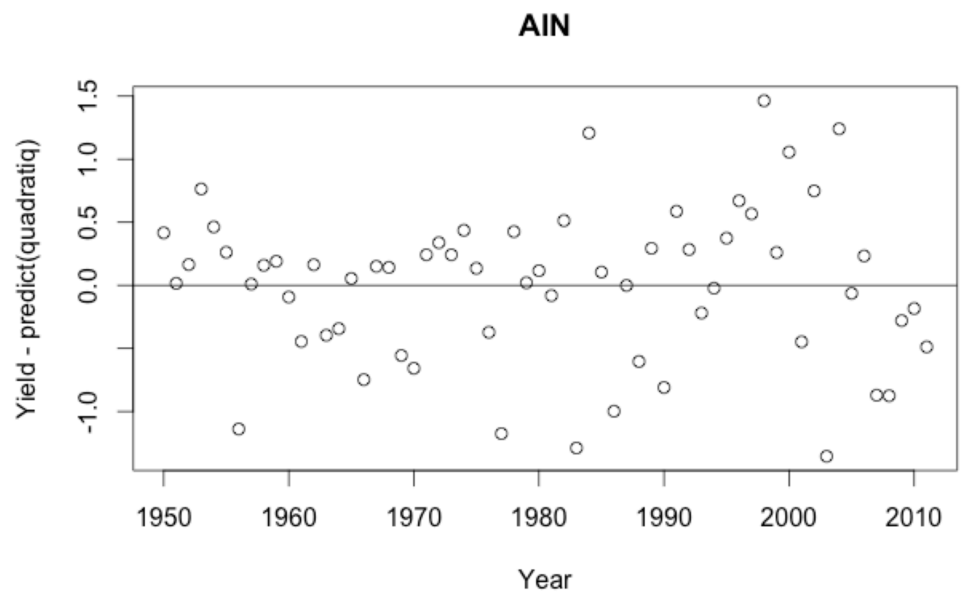
```
> print(summary( lplateau))

Formula: Yield ~ LP(Year, Ymax, Tmax, P)

Parameters:
      Estimate Std. Error t value Pr(>|t|)
Ymax 6.778e+00  2.014e-01   33.66   <2e-16 ***
P    1.091e-01  5.424e-03   20.11   <2e-16 ***
Tmax 2.002e+03  2.382e+00  840.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05

Residual standard error: 0.6041 on 59 degrees of

Number of iterations to convergence: 3
Achieved convergence tolerance: 4.608e-09

> print(AIC(lplateau))
[1] 118.3727
```

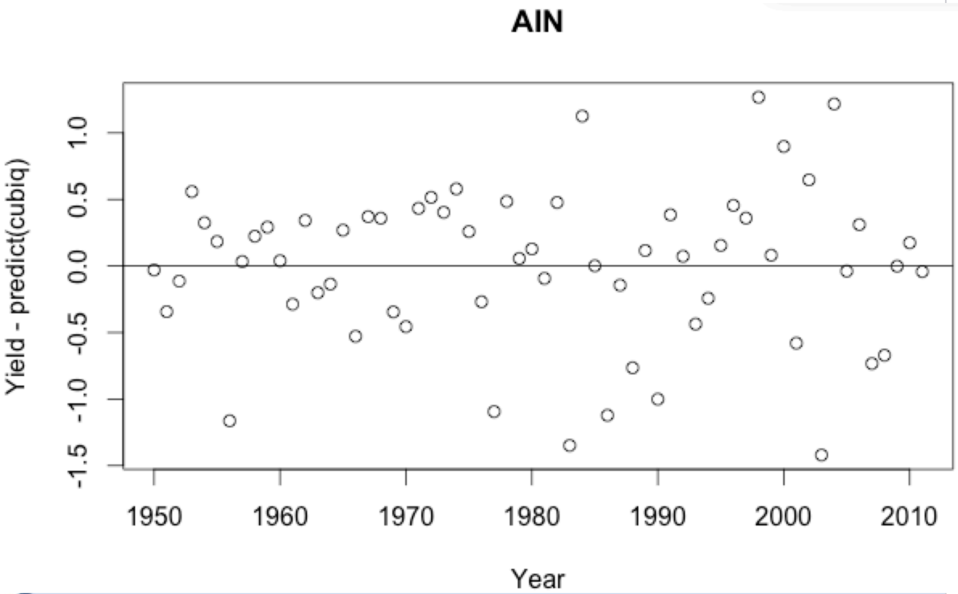
AIN

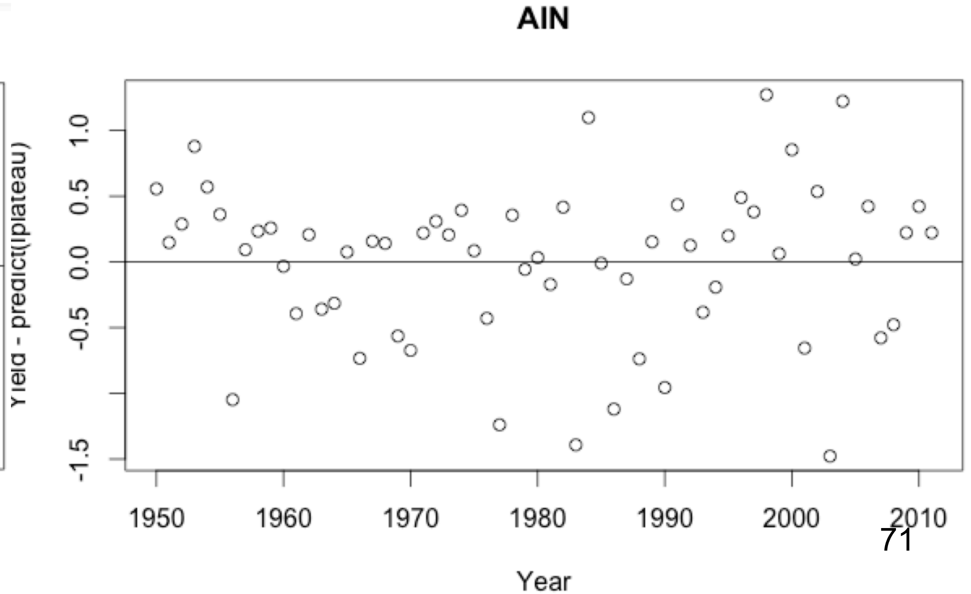
```
> ########################
> ##    Calcul de R2    ##
> ########################
>
> #R2 pour le modele lineaire
> R_lineaire <-1-((sum((Yield-predict(lineaire))^2))/(sum((Yield-mean(Yield))^2)))
> print(R_lineaire)
[1] 0.9018137
>
> #R2 pour le modele quadratique
> R_quadratiq <-1-((sum((Yield-predict(quadratiq))^2))/(sum((Yield-mean(Yield))^2))
> print(R_quadratiq)
[1] 0.9018334
>
> #R2 pour le modele cubique
> R_cubiq <-1-((sum((Yield-predict(cubiq))^2))/(sum((Yield-mean(Yield))^2)))
> print(R_cubiq)
[1] 0.9110842
>
> #R2 pour le modele lineaire + plateau
> R_lplateau <-1-((sum((Yield-predict(lplateau))^2))/(sum((Yield-mean(Yield))^2)))
> print(R_lplateau)
[1] 0.9069815
```

AIN


AIN
Yield - predict(lineaire)
Year
1950
1960
1970
1980
1990
2000
2010
-1.0
0.0
0.5
1.0
1.5


AIN
Yield - predict(quadratiq)
Year
1950
1960
1970
1980
1990
2000
2010
-1.0
0.0
0.5
1.0
1.5

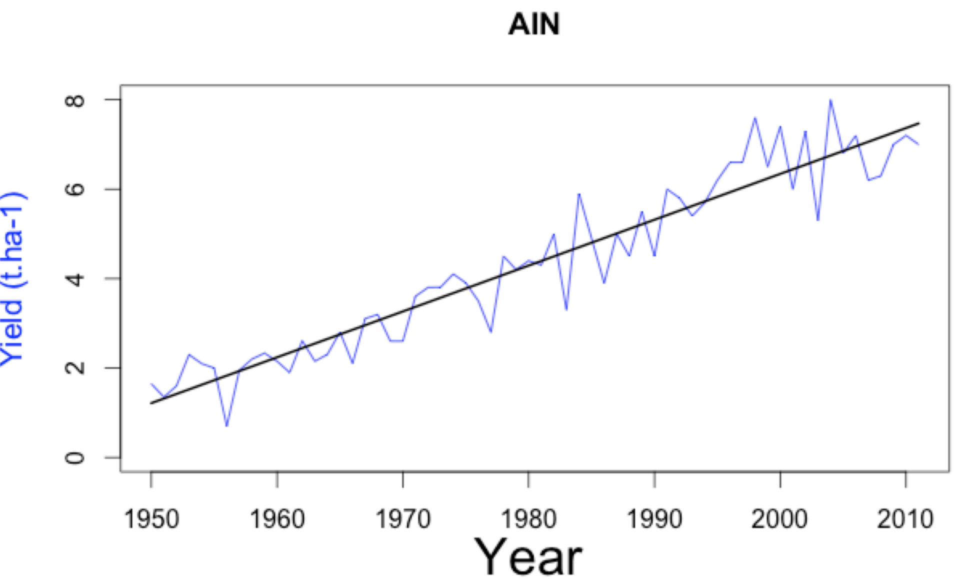

AIN
Yield - predict(cubiq)
Year
1950
1960
1970
1980
1990
2000
2010
-1.5
-1.0
-0.5
0.0
0.5
1.0

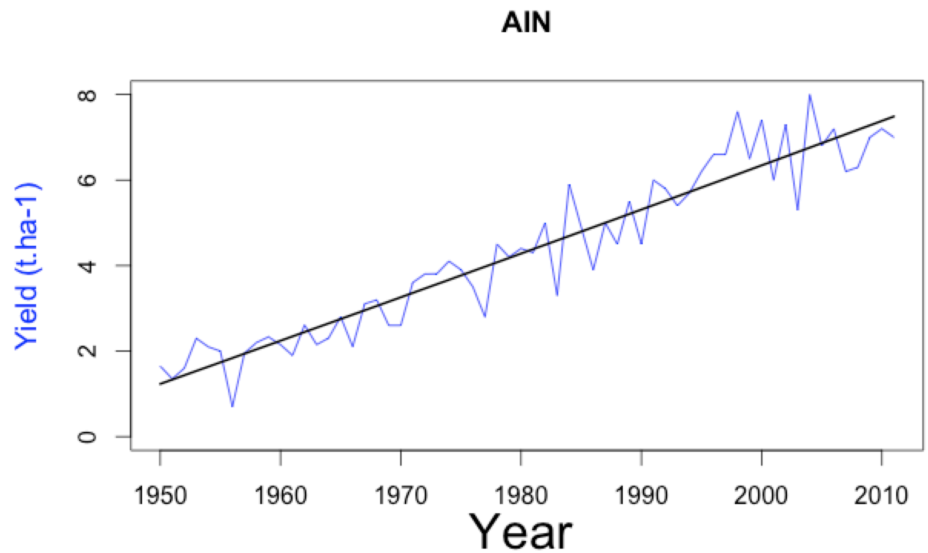

AIN
Yield - predict(plateau)
Year
1950
1960
1970
1980
1990
2000
2010
-1.5
-0.5
0.0
0.5
1.0

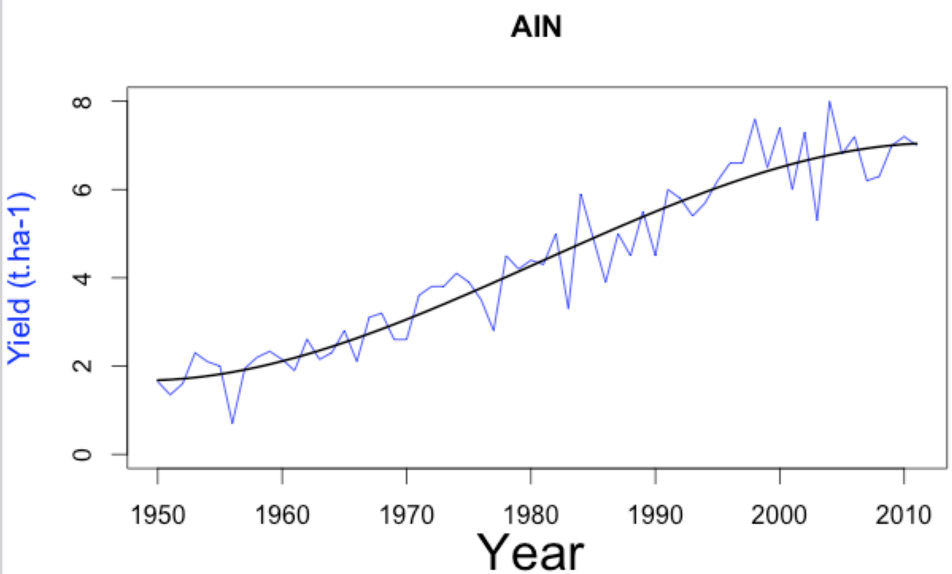
AIN

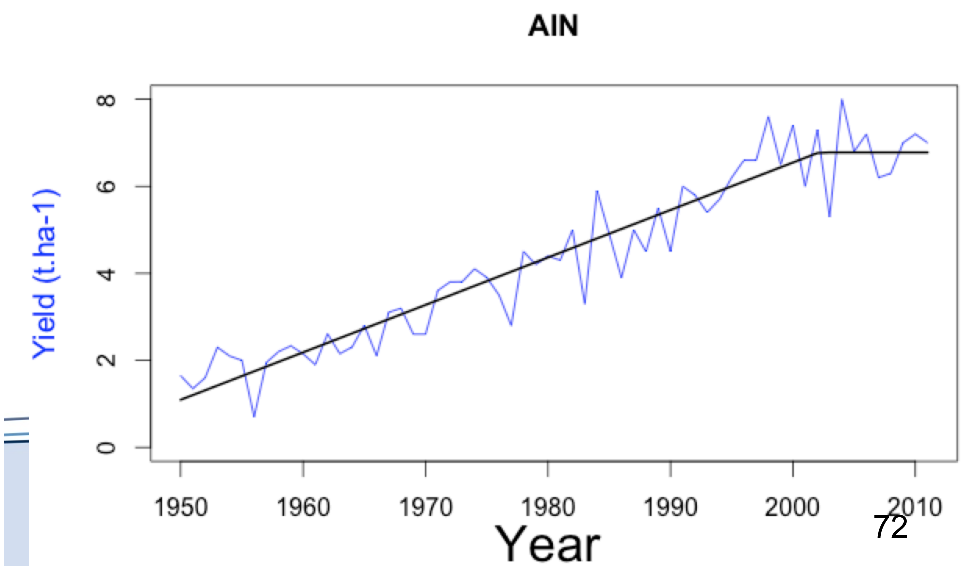
Le réseau des instituts des filières animales et végétales

SCIENCE & IMPACT









AIN

	Linéaire	Quadratique	Cubique	Linéaire + plateau
R^2	0,9018	0,9018	**0,9111**	0,907
AIC	119,72	121,71	**117,58**	118,37

EURE

```
> print(summary(lineaire))

Call:
glm(formula = Yield ~ Year)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4516  -0.3786   0.1103   0.3037   1.5265

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -204.00754    8.33922  -24.46   <2e-16 ***
Year           0.10599    0.00421   25.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.

(Dispersion parameter for gaussian family taken to be 0

    Null deviance: 244.17  on 61  degrees of freedom
Residual deviance:  21.12  on 60  degrees of freedom
AIC: 115.18

Number of Fisher Scoring iterations: 2
```

```
> print(summary(quadratiq))

Call:
glm(formula = Yield ~ Year_b + Year2_b)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.61785  -0.39234   0.05453   0.40102   1.35581

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.3319058  0.2124793   10.97 7.03e-16 ***
Year_b       0.1397896  0.0161053    8.68 3.92e-12 ***
Year2_b     -0.0005541  0.0002554   -2.17   0.0341 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.

(Dispersion parameter for gaussian family taken to be 0

    Null deviance: 244.168  on 61  degrees of freedom
Residual deviance:  19.559  on 59  degrees of freedom
AIC: 112.42

Number of Fisher Scoring iterations: 2
```

EURE

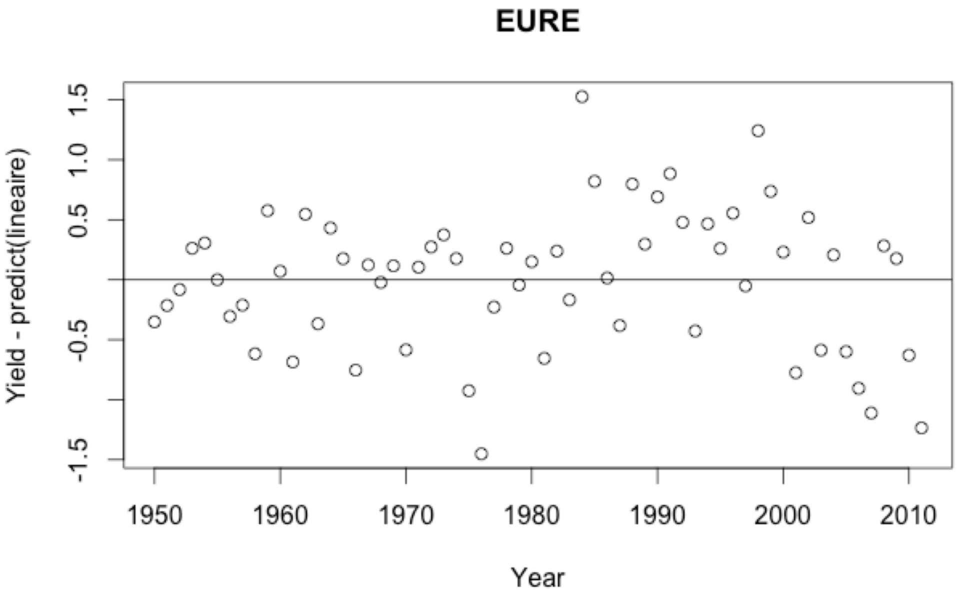
```
> print(summary(cubiq))

Call:
glm(formula = Yield ~ Year_b + Year2_b + Year3_b)

Deviance Residuals:
     Min        1Q     Median       3Q       Max
-1.50835  -0.41477   0.07588  0.35000   1.26941

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.804e+00  2.604e-01  10.770 1.85e-15 ***
Year_b       4.287e-02  3.727e-02   1.150  0.25482
Year2_b      3.451e-03  1.426e-03   2.419  0.01872 *
Year3_b     -4.377e-05  1.536e-05  -2.849  0.00607 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.

(Dispersion parameter for gaussian family taken to be 0

    Null deviance: 244.168  on 61  degrees of freedom
Residual deviance:  17.159  on 58  degrees of freedom
AIC: 106.3

Number of Fisher Scoring iterations: 2
```

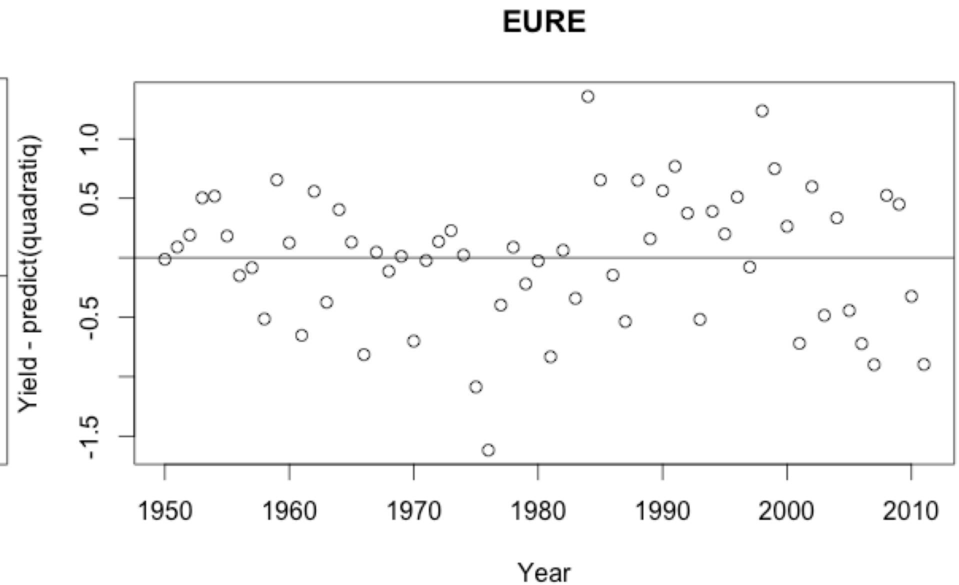
```
> #Modele lineaire+plateau
> lplateau <- nls(Yield ~LP(Year, Ymax, Tmax, P),sta
23.3341 :      8.50     0.12 1998.00
16.90515 :      8.215385     0.119213 1998.544287
16.90514 :      8.215385     0.119213 1998.547880
> print(summary( lplateau))

Formula: Yield ~ LP(Year, Ymax, Tmax, P)

Parameters:
      Estimate Std. Error t value Pr(>|t|)
Ymax 8.215e+00  1.485e-01   55.34   <2e-16 ***
P    1.192e-01  5.407e-03   22.05   <2e-16 ***
Tmax 1.999e+03  1.789e+00 1116.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.

Residual standard error: 0.5353 on 59 degrees of fre

Number of iterations to convergence: 2
Achieved convergence tolerance: 3.037e-09

> print(AIC(lplateau))
[1] 103.3783
```

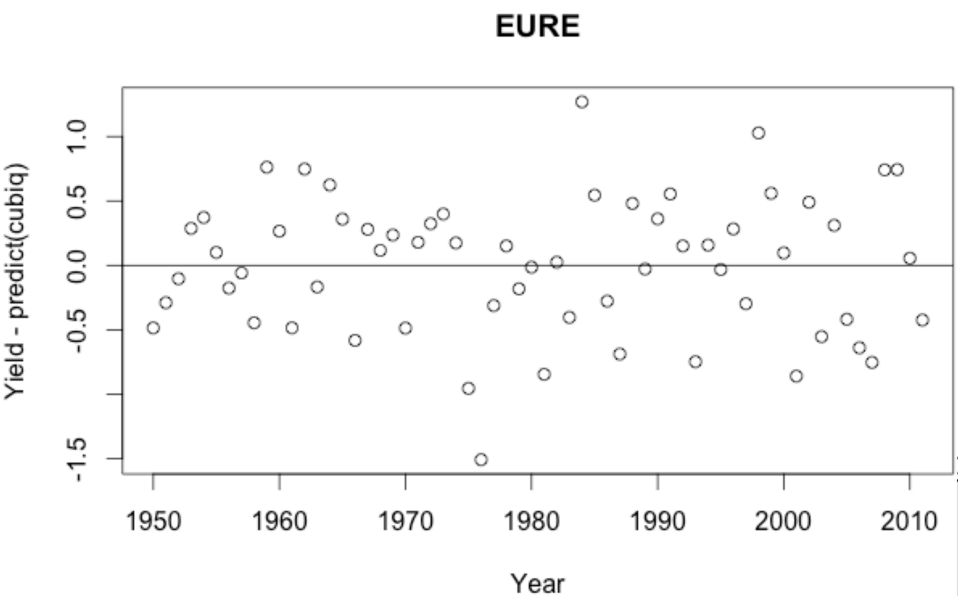
EURE

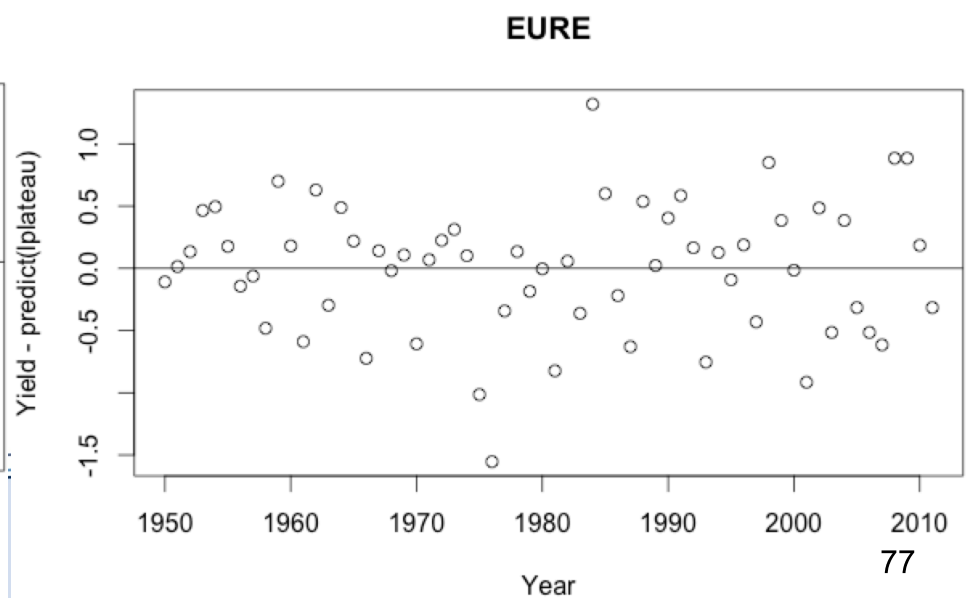
```
>
> ########################
> ##    Calcul de R2    ##
> ########################
>
> #R2 pour le modele lineaire
> R_lineaire <-1-((sum((Yield-predict(lineaire))^2))/(sum((Yield-mean(Yield))^2)))
> print(R_lineaire)
[1] 0.9135014
>
> #R2 pour le modele quadratique
> R_quadratiq <-1-((sum((Yield-predict(quadratiq))^2))/(sum((Yield-mean(Yield))^2)))
> print(R_quadratiq)
[1] 0.9198934
>
> #R2 pour le modele cubique
> R_cubiq <-1-((sum((Yield-predict(cubiq))^2))/(sum((Yield-mean(Yield))^2)))
> print(R_cubiq)
[1] 0.9297247
>
> #R2 pour le modele lineaire + plateau
> R_lplateau <-1-((sum((Yield-predict(lplateau))^2))/(sum((Yield-mean(Yield))^2)))
> print(R_lplateau)
[1] 0.9307643
>
```

EURE

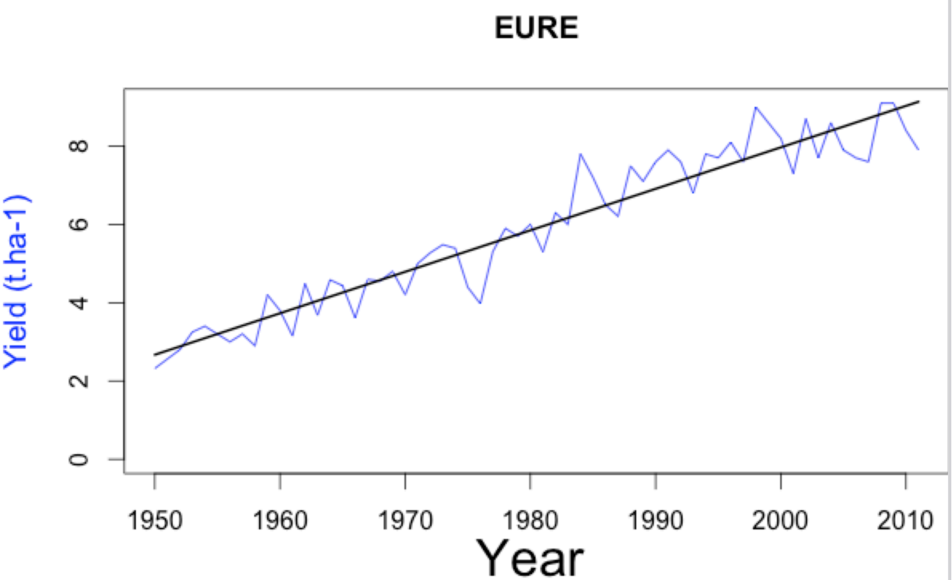

EURE
Yield - predict(lineaire)
-1.5
-0.5
0.5
1.0
1.5
1950
1960
1970
1980
1990
2000
2010
Year

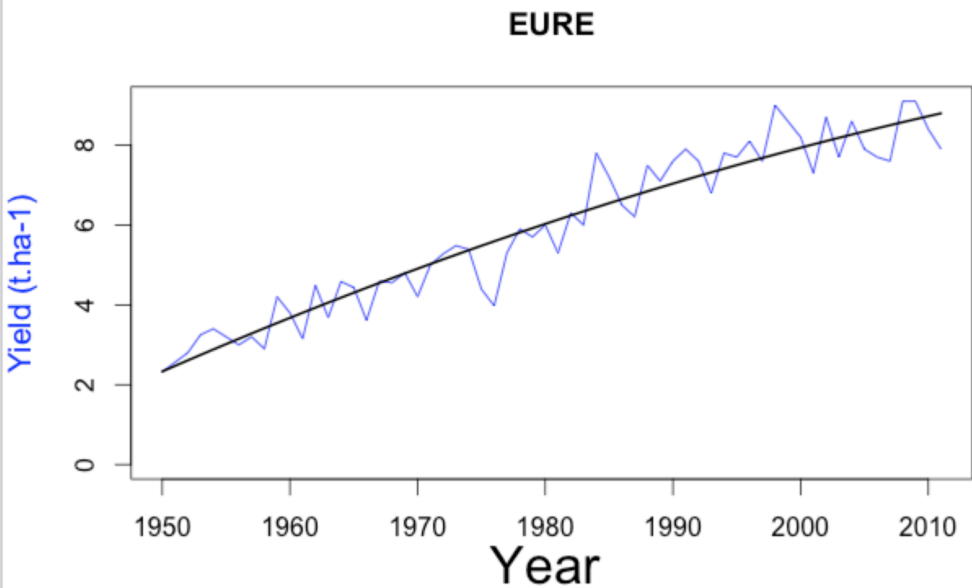

EURE
Yield - predict(quadratiq)
-1.5
-0.5
0.5
1.0
1950
1960
1970
1980
1990
2000
2010
Year

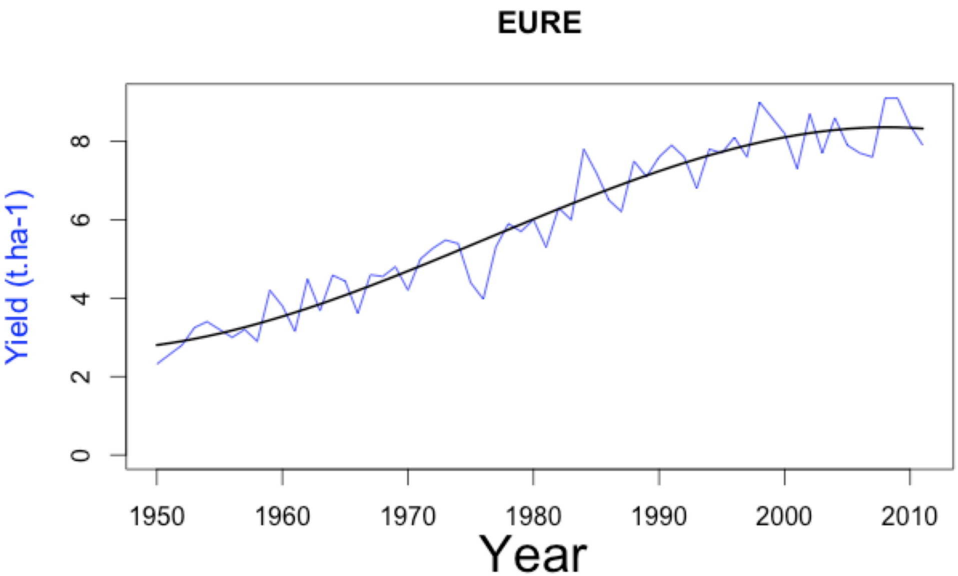

EURE
Yield - predict(cubiq)
-1.5
-0.5
0.0
0.5
1.0
1950
1960
1970
1980
1990
2000
2010
Year

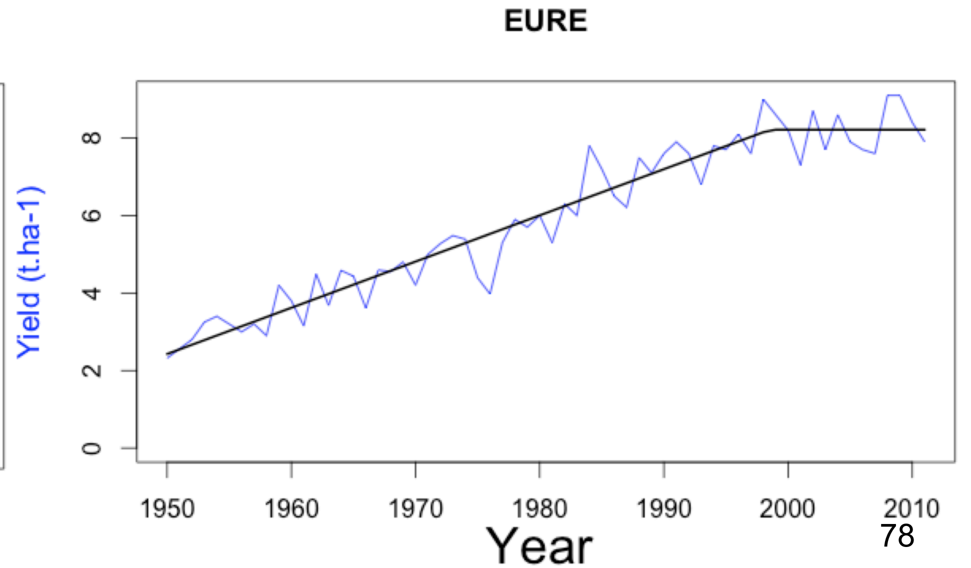

EURE
Yield - predict(lplateau)
-1.5
-0.5
0.0
0.5
1.0
1950
1960
1970
1980
1990
2000
2010
Year

EURE


EURE
Yield (t.ha-1)
0
2
4
6
8
1950
1960
1970
1980
1990
2000
2010
Year


EURE
Yield (t.ha-1)
0
2
4
6
8
1950
1960
1970
1980
1990
2000
2010
Year


EURE
Yield (t.ha-1)
0
2
4
6
8
1950
1960
1970
1980
1990
2000
2010
Year


EURE
Yield (t.ha-1)
0
2
4
6
8
1950
1960
1970
1980
1990
2000
2010
Year

EURE

	Linéaire	Quadratique	Cubique	Linéaire + plateau
R^2	0,9135	0,9199	0,9297	**0,9308**
AIC	115,18	112,42	106,3	**103,38**

CREUSE

```
> print(summary(lineaire))

Call:
glm(formula = Yield ~ Year)

Deviance Residuals:
     Min        1Q    Median       3Q      Max
-1.20085  -0.31264  0.04427  0.36971  1.04112

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.466e+02  7.321e+00  -20.03   <2e-16 ***
Year         7.580e-02  3.696e-03   20.51   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.

(Dispersion parameter for gaussian family taken to be 0

    Null deviance: 130.367  on 61  degrees of freedom
Residual deviance:  16.277  on 60  degrees of freedom
AIC: 99.03

Number of Fisher Scoring iterations: 2
```

```
> print(summary(quadratiq))

Call:
glm(formula = Yield ~ Year_b + Year2_b)

Deviance Residuals:
     Min        1Q    Median       3Q      Max
-1.21214  -0.30756  0.05393  0.36087  1.03801

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.148e+00  1.938e-01   5.925 1.72e-07 ***
Year_b       7.818e-02  1.469e-02   5.322 1.67e-06 ***
Year2_b     -3.892e-05  2.329e-04  -0.167    0.868
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.

(Dispersion parameter for gaussian family taken to be 0

    Null deviance: 130.367  on 61  degrees of freedom
Residual deviance:  16.269  on 59  degrees of freedom
AIC: 101

Number of Fisher Scoring iterations: 2
```

CREUSE

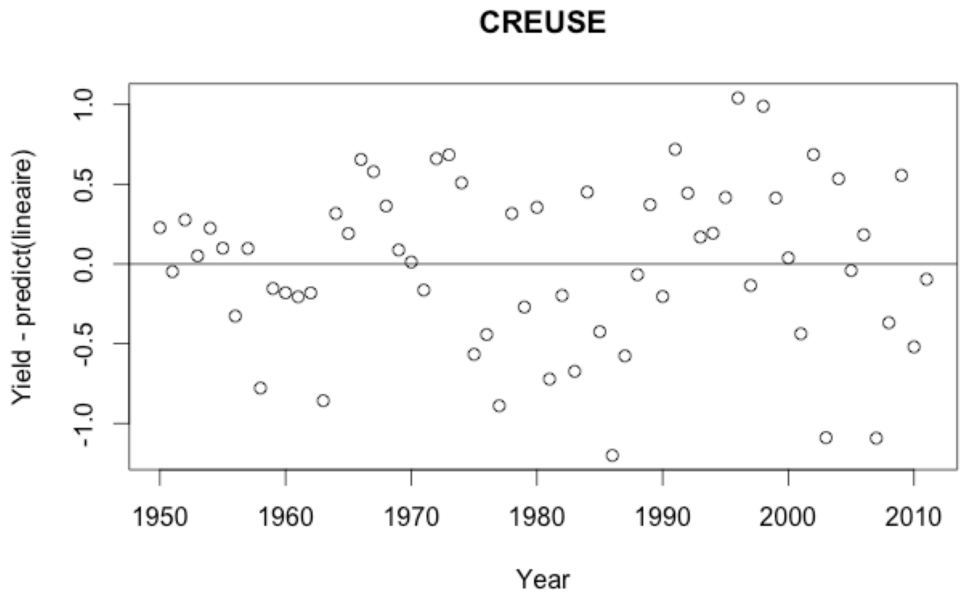
```
> #Model cubique
> cubiq <- glm(Yield~Year_b+Year2_b+Year3_b)
> print(summary(cubiq))

Call:
glm(formula = Yield ~ Year_b + Year2_b + Year3_b)

Deviance Residuals:
     Min         1Q     Median         3Q        Max
-1.25814   -0.28173    0.08025    0.34792    0.95823

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.314e+00  2.513e-01   5.228 2.45e-06 ***
Year_b       4.425e-02  3.596e-02   1.230    0.223
Year2_b      1.363e-03  1.376e-03   0.990    0.326
Year3_b     -1.532e-05  1.483e-05  -1.033    0.306
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.

(Dispersion parameter for gaussian family taken to be 0

    Null deviance: 130.367  on 61  degrees of freedom
Residual deviance:  15.975  on 58  degrees of freedom
AIC: 101.87

Number of Fisher Scoring iterations: 2
```

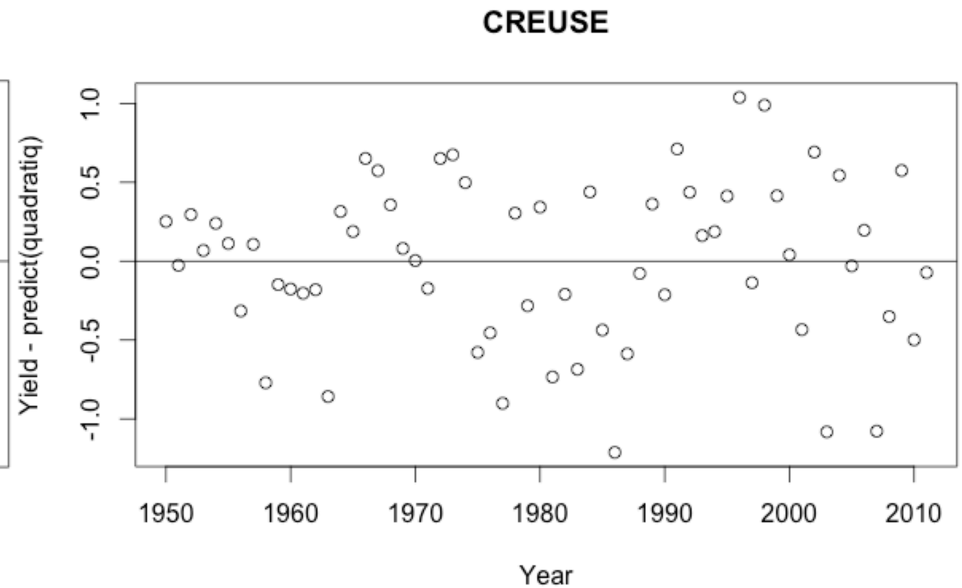
```
> print(summary( lplateau))

Formula: Yield ~ LP(Year, Ymax, Tmax, P)

Parameters:
      Estimate Std. Error t value Pr(>|t|)
Ymax 5.278e+00  1.740e-01   30.33   <2e-16 ***
P    8.052e-02  4.688e-03   17.18   <2e-16 ***
Tmax 2.002e+03  2.787e+00  718.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05

Residual standard error: 0.522 on 59 degrees of

Number of iterations to convergence: 2
Achieved convergence tolerance: 9.916e-06

> print(AIC(lplateau))
[1] 100.2729
```

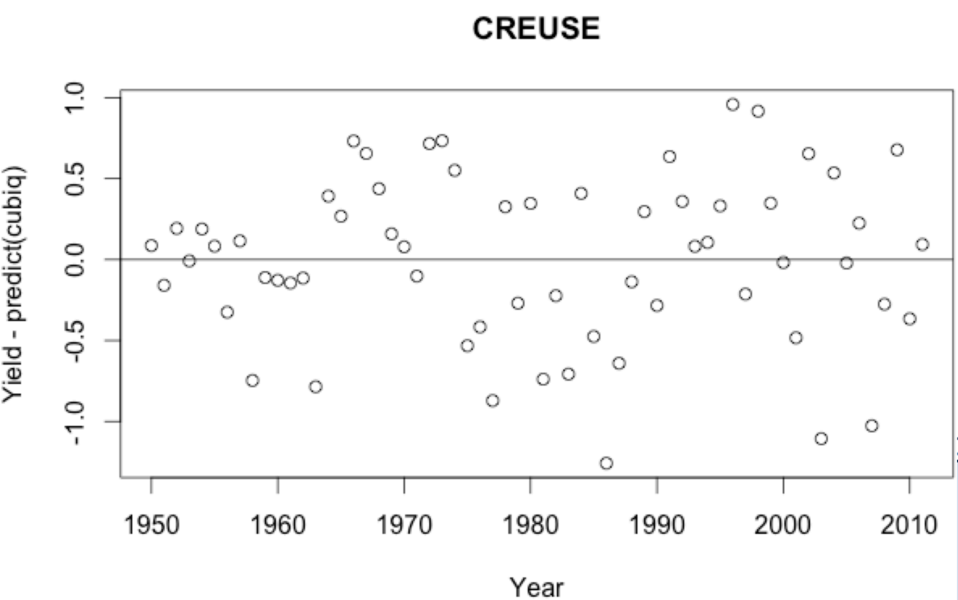
CREUSE

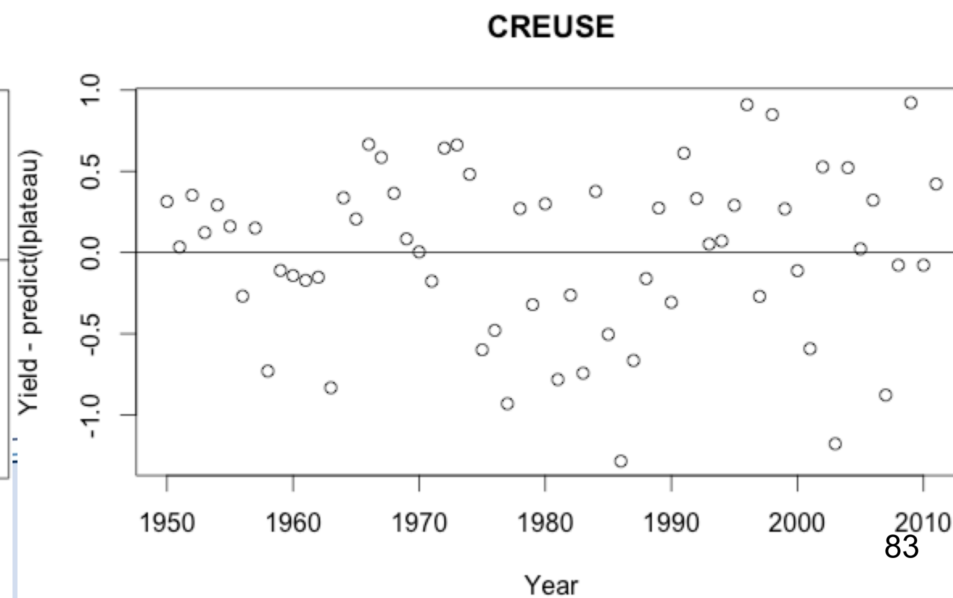
```
> ########################
> ##    Calcul de R2    ##
> ########################
>
> #R2 pour le modele lineaire
> R_lineaire <-1-((sum((Yield-predict(lineaire))^2))/(sum((Yield-mean(Yield))^2)))
> print(R_lineaire)
[1] 0.8751465
>
> #R2 pour le modele quadratique
> R_quadratiq <-1-((sum((Yield-predict(quadratiq))^2))/(sum((Yield-mean(Yield))^2)))
> print(R_quadratiq)
[1] 0.8752056
>
> #R2 pour le modele cubique
> R_cubiq <-1-((sum((Yield-predict(cubiq))^2))/(sum((Yield-mean(Yield))^2)))
> print(R_cubiq)
[1] 0.8774615
>
> #R2 pour le modele lineaire + plateau
> R_lplateau <-1-((sum((Yield-predict(lplateau))^2))/(sum((Yield-mean(Yield))^2)))
> print(R_lplateau)
[1] 0.8766617
```

CREUSE


CREUSE
Yield - predict(lineaire)
Year
1950
1960
1970
1980
1990
2000
2010
-1.0
-0.5
0.0
0.5
1.0

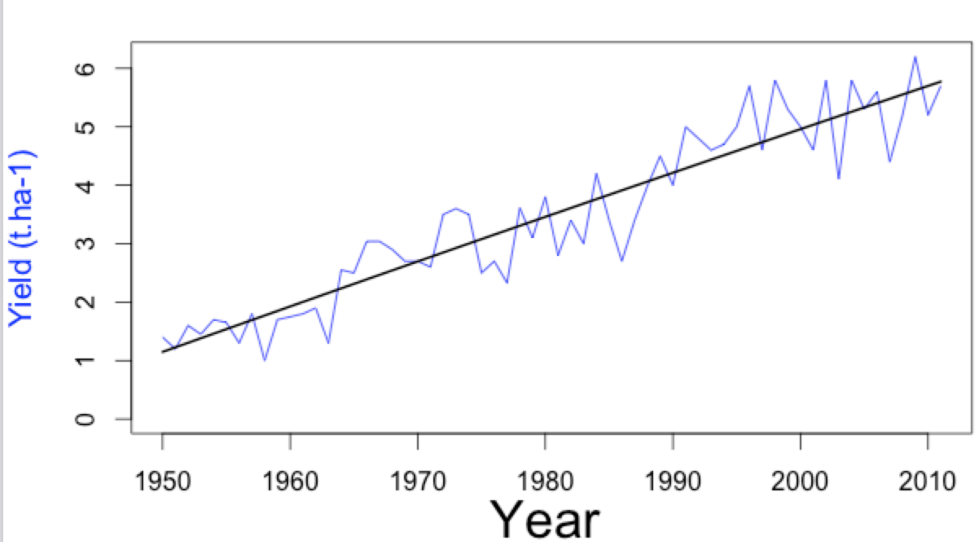

CREUSE
Yield - predict(quadratiq)
Year
1950
1960
1970
1980
1990
2000
2010
-1.0
-0.5
0.0
0.5
1.0


CREUSE
Yield - predict(cubiq)
Year
1950
1960
1970
1980
1990
2000
2010
-1.0
-0.5
0.0
0.5
1.0

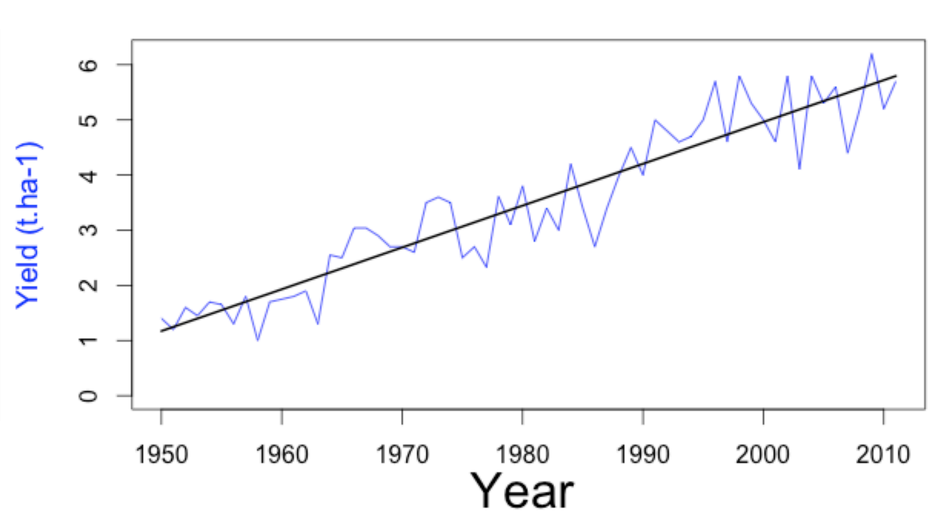

CREUSE
Yield - predict(lplateau)
Year
1950
1960
1970
1980
1990
2000
2010
-1.0
-0.5
0.0
0.5
1.0

CREUSE

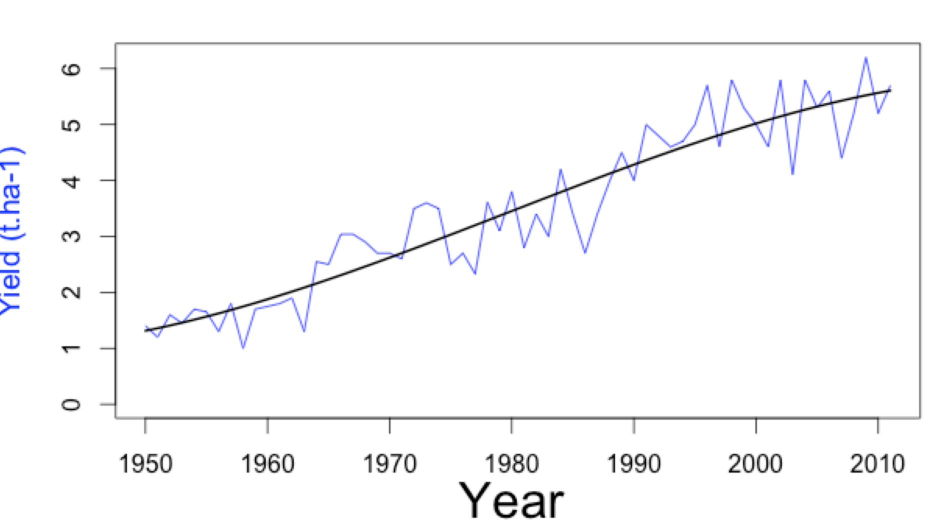
CREUSE



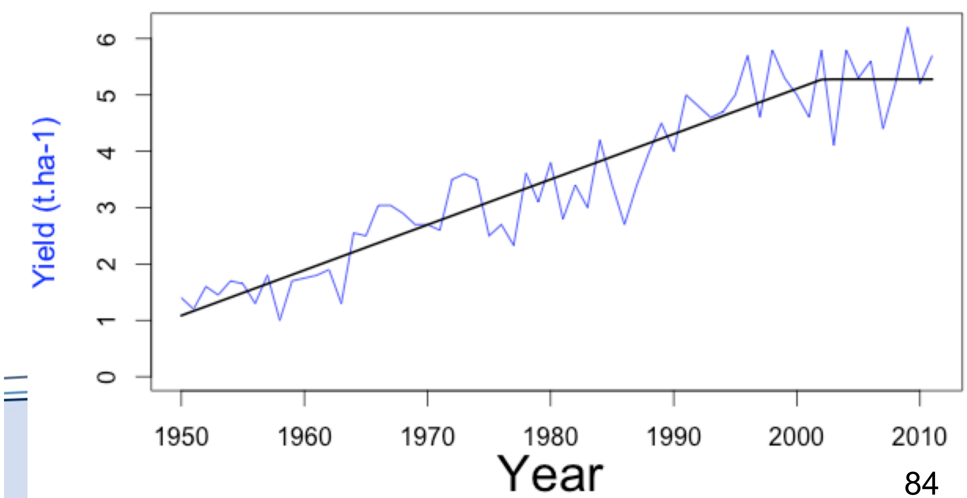
CREUSE



CREUSE



CREUSE



CREUSE

	Linéaire	Quadratique	Cubique	Linéaire + plateau
R^2	0,8751	0,8752	**0,8775**	0,8767
AIC	**99,03**	101	101,87	100.27