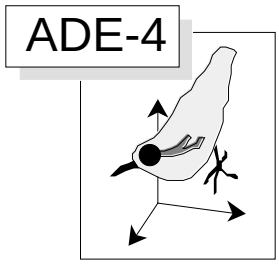


Variables explicatives indépendantes

Résumé

La fiche décrit la régression linéaire dans le cas le plus simple et le plus clair, celui de variables explicatives indépendantes. La première situation abordée est celle d'une variable explicative unique (Module UniVar). On considère ensuite k explicatives indépendantes (Module OrthoVar). La régression sur composantes principales (PCR) et la régression sur vecteurs propres de voisinages (NER) sont de ce type. On souligne que le principe de régression linéaire permet d'aborder des structures qui ne le sont pas du tout comme des courbes de réponses en cloche ou des chroniques auto corrélées négativement.

Plan

D. Chessel et J. Thioulouse

1 — Introduction : méthodes et modèles linéaires

Les travaux de C.J.F. Ter Braak pose avec la plus grande précision la question des modèles écologiques. Ils sont de deux sortes. Soit il s'agit de modéliser l'impact de l'environnement sur l'abondance et la distribution inter-spécifiques des organismes vivants (*response modelling*), soit il s'agit de prédire des paramètres environnementaux grâce à l'abondance ou la distribution inter-spécifiques des organismes (*Calibration*). Les deux questions mettent en œuvre un tableau floro-faunistique X et un tableau de variables environnementales Y portant sur les mêmes stations, les lignes communes aux deux tableaux.

Si des centaines d'articles traitent de ces questions, sans pour autant qu'aucune méthode ne se soit imposée, c'est que tant au plan écologique que statistique la question est complexe. Les deux questions sont très différentes. Elles permettent de distinguer deux familles pourtant étroitement liées de méthodes statistiques.

La première fait des variables mésologiques des variables explicatives et des variables faunistiques des variables à expliquer. Les taxons étudiés sont tellement nombreux, en général, que modéliser les courbes de réponses espèce par espèce pose immanquablement la question des modèles de synthèse. D'autant plus, que le partage de l'espace entre niches écologiques implique que l'abondance d'une espèce implique grossièrement que l'abondance des autres diminuent. Si il y a plusieurs variables à expliquer, la première question concerne donc le choix entre des modèles séparés, des modèles communs, des modèles concurrentiels.

La seconde fait des variables faunistiques des variables explicatives et des variables mésologiques des variables à expliquer. Les variables à prédire sont en petit nombre, et même en général uniques, mais les taxons étudiés sont tellement nombreux que modéliser pose immanquablement la question du nombre de variables. Inférer un milieu à partir d'un cortège floro-faunistique (*calibration*) doit tenir compte, d'après Ter Braak (1993[1]), des faits que :

1 - le nombre d'espèces est grand (10-300) et la multicolinéarité des prédicteurs assurée ;

2 - les données contiennent beaucoup de valeurs nulles et le total par site est sans signification (ce qui est discutable dès qu'on traite de la pollution) ;

3 - les relations sont non linéaires à cause de la loi de Shelford[2] et de la séparation des niches (Whittaker et Coll. 1973[3]).

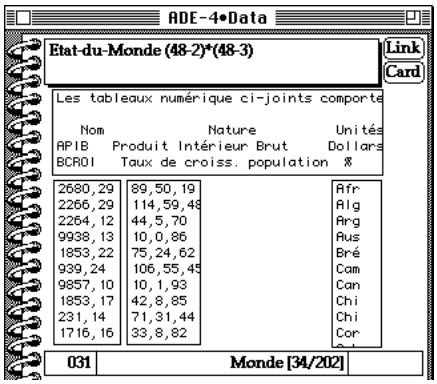
L'essentiel des problèmes posés est abordé. Que faire avec de très nombreuses variables explicatives ? Que faire avec des courbes de réponses non linéaires ? Que faire dans un sens (milieu fonction de la faune) quand on sait des choses dans l'autre sens (l'abondance fonction de milieu est non linéaire) ? Que faire avec de nombreuses variables à expliquer ? Que faire des variables explicatives très corrélées ?

Il va sans dire que choisir une méthode statistique de prédiction nécessite dans la plupart des cas une vue d'ensemble des méthodes de régression qui comprend au moins les régressions sur variables indépendantes et en particulier les régressions sur composantes principales, les régressions multiples classiques et les régressions pas à pas, les régressions PLS, les régressions locales, les régressions par boules, les analyses sur variables instrumentales. Il convient pour le moins d'avoir une idée des pièges principaux. Le présent fascicule apporte des éléments de discussion et la description de certains outils dans ce domaine particulièrement difficile de la modélisation statistique.

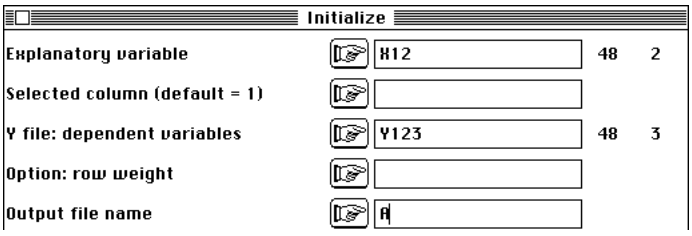
Il convient d'abord de distinguer entre méthodes linéaires et modèles linéaires. Ce n'est pas parce qu'on utilise des principes géométriques simples (méthodes) qu'on obtient forcément des modèles simples apparentés à la seule droite de régression.

2 — Régression univariée

La situation abordée est celle d'une variable explicative mesurée sur n échantillons (ou individus, lignes, relevés...) qui doit prédire k variables à expliquer mesurées sur les mêmes échantillons. On ne s'intéresse qu'à une prédiction séparées de l'unique explicative sur chacune des expliquées (il peut y avoir, bien sûr qu'une variable à expliquer). La première question est celle du modèle linéaire ou non linéaire au sens commun du terme. Les expliquées sont elles des fonctions simples (y = ax + b) de la variable prédictrice ?



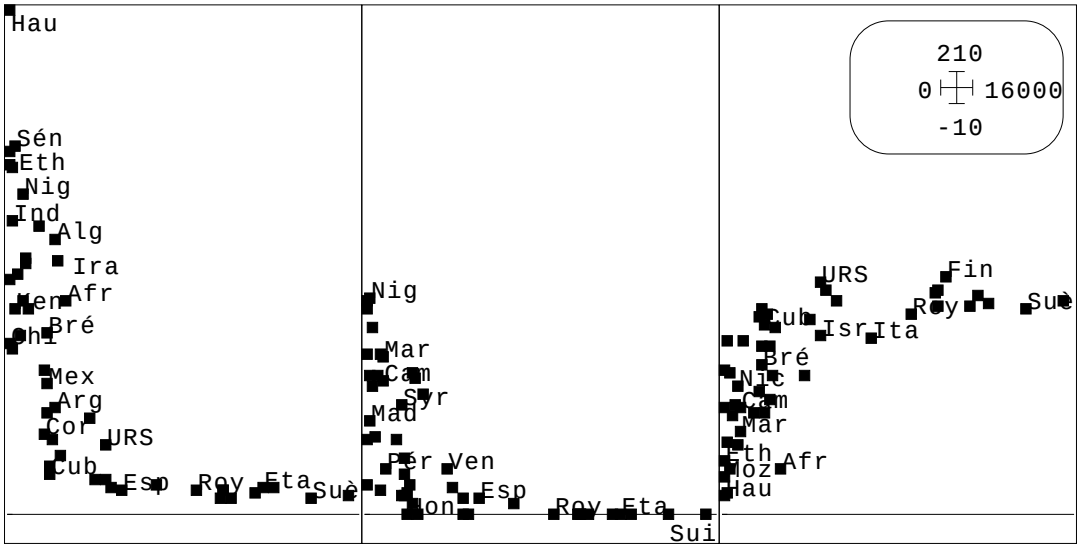
Utiliser la carte Monde de la pile ADE-4•Data. Le champ de gauche donne un fichier binaire X12 (48 lignes et 2 colonnes). Celui du milieu donne un fichier Y123 (48 lignes et 3 colonnes). Les tableaux numérique ci-joints comporte 49 lignes et respectivement 2 et 3 colonnes. Ces données brutes font partie d'un ensemble de statistiques publiées dans "L'état du Monde 1984" (Édition La Découverte). La plupart des valeurs concernent 1983, certaines sont associées au dernier recensement de chaque pays. La première variable de X12 est le PIB (Produit Intérieur Brut) par habitant exprimée en dollars. Les observations pour les pays à économie planifiée sont des estimations C.I.A. La seconde variable de X12 est le taux de croissance de la population exprimé en 1 pour 1000. Ces deux variables seront dites explicatives et forment le premier tableau. La première variable de Y123 est le taux (en 1/1000) de mortalité infantile, nombre de décès d'enfants âgés de moins d'un an rapporté au nombre d'enfants nés vivants pendant l'année étudiée. La seconde variable est le taux (%) d'analphabétisme, soit la proportion des illettrés dans la population de plus de quinze ans. La troisième variable est le taux (%) d'inscription scolaire pour la catégorie (approximative suivant les pays) des 11-17 ans. Les variables de Y123 sont dites à expliquer. On ne s'occupe maintenant que de la prédiction des 3 variables de Y123 par la première variable de X12. Ouvrir le module UniVarReg :



On obtient :

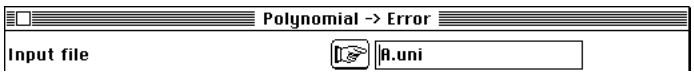
```
New TEXT file A.uni contains the parameters:
----> Explanatory variables: X12 [48][2]
----> Selected variable: 1
----> Dependant variable file: Y123 [48][3]
----> Row weighting file: Uniform_weighting
-------------------------------------------------------
```

Dans Curves, représenter les données brutes :



2.1 — Régressions polynomiales

Les relations ne sont manifestement pas linéaires. Les méthodes pour pallier à cet inconvénient le sont néanmoins. La première est la régression polynomiale. Pour choisir le degré du polynôme utiliser l'option :



On a :

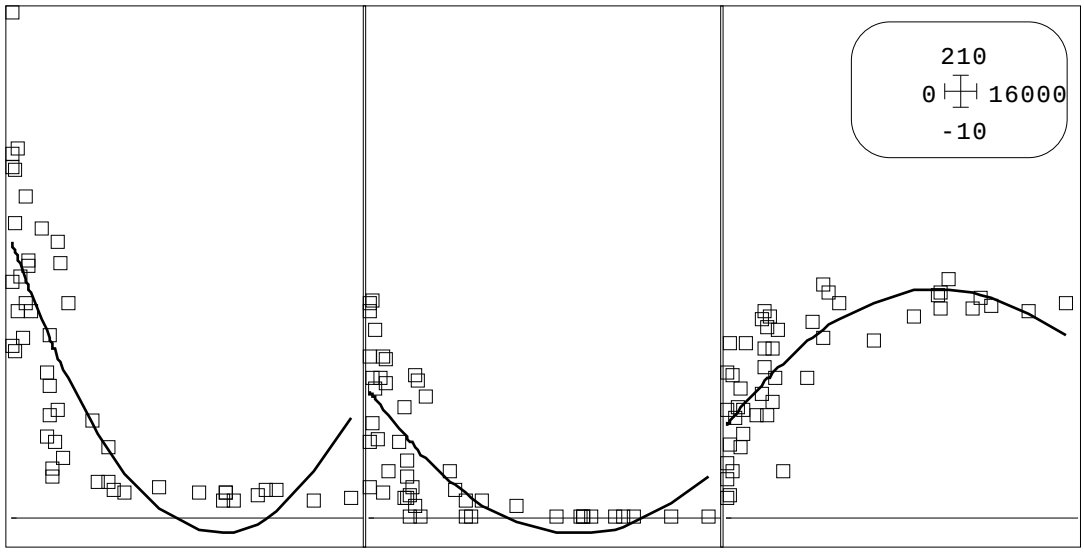
```
Polynomial model
Y file: Y123
--- Number of rows: 48, columns: 3
X coordinate file: X12
--- Number of rows: 48, columns: 2
--- Selected column: 1
------------------------------------------------------------------------
Predicted variable n° 1
------------------------------------------------------------------------
 d°|Error/Vari|    a0    |    a1    |    a2    |    a3    |    a4    |
------------------------------------------------------------------------
  1| 5.120e-01| 6.213e+01|-3.573e+01|
  2| 3.493e-01| 3.831e+01|-6.522e+01| 2.382e+01|
  3| 3.040e-01| 2.655e+01|-6.031e+01| 4.934e+01|-1.112e+01|
  4| 3.008e-01| 2.727e+01|-4.927e+01| 5.153e+01|-2.296e+01| 3.577e+00|
  5| 3.004e-01| 2.847e+01|-4.631e+01| 4.598e+01|-2.678e+01| 9.249e+00|
  6| 3.002e-01| 3.109e+01|-4.331e+01| 2.920e+01|-2.554e+01| 2.625e+01|•••
------------------------------------------------------------------------
```

L'information est donnée pour chaque variable. L'erreur est exprimée en pourcentage de la variance (laquelle est ici l'erreur minimum associée au modèle trivial y = constante). On lit par exemple que l'erreur commise par une régression polynomiale de degré 2 vaut 34.9 % de la variance et que le meilleur polynôme de degré 2 s'écrit :

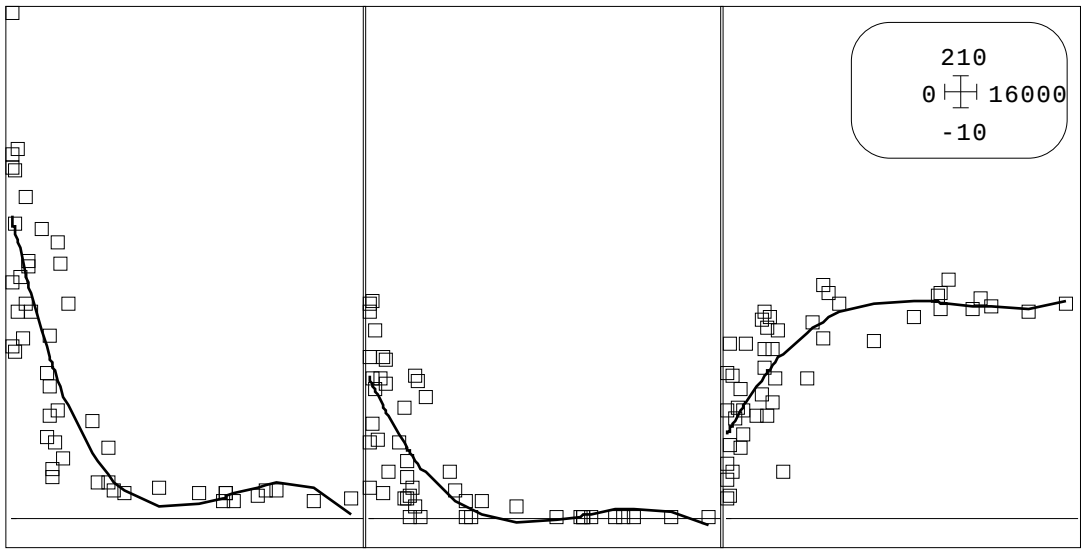
$$y = 38.31 - 65.22\ x + 23.82\ x^2.$$

Noter que ces coefficients portent sur la variable x centrée et réduite.

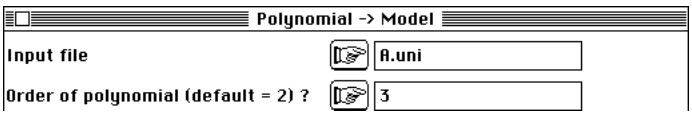
Ceci permet de comparer les équations d'une variable prédite à l'autre. Pour représenter graphiquement le résultat de ce modèle on peut utiliser directement CurveModels :



Avec un polynôme de degré 3, on obtient :



On peut aussi récupérer les prédictions polynomiales dans UniVarReg :

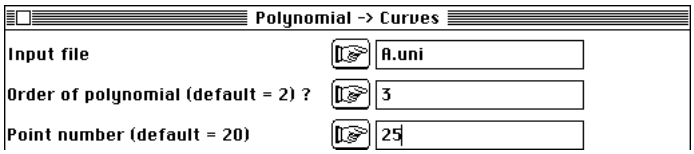


On a alors :

```
Polynomial model
Y file: Y123
--- Number of rows: 48, columns: 3
X coordinate file: X12
--- Number of rows: 48, columns: 2
--- Selected column: 1
Order of polynomial: 3
Output file: Ad°3
--- Number of rows: 48, columns: 3
```

```
This file contains the polynomial model for each variable (columns),
```

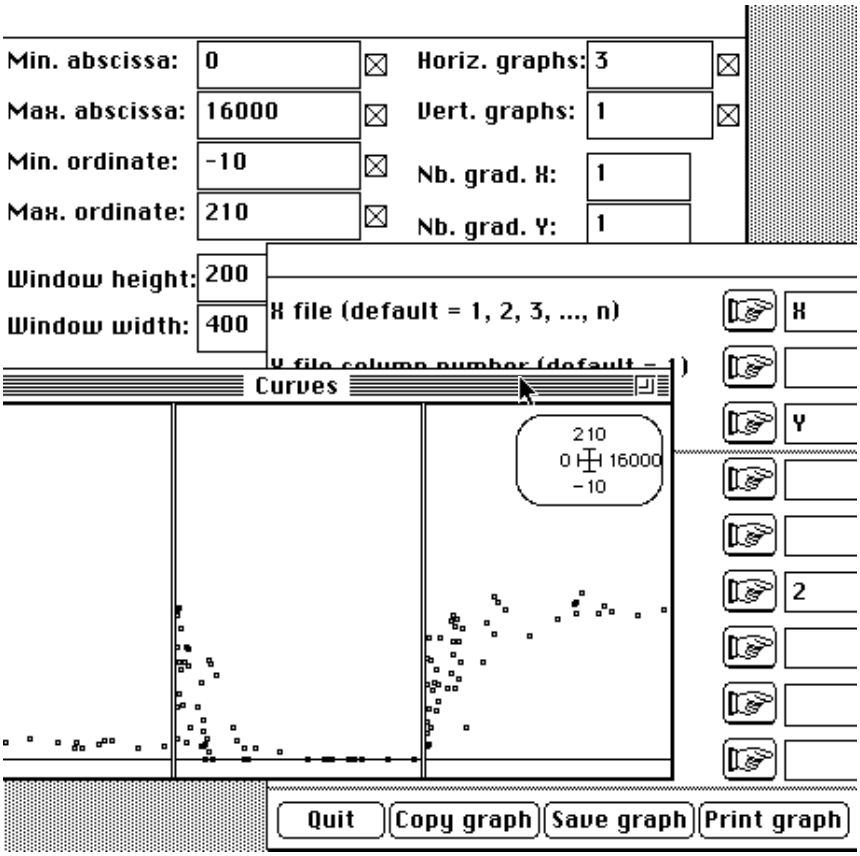
On obtient les prévisions pour les points de mesures. On peut les obtenir pour des valeurs régulièrement réparties sur l'intervalle de variation par :



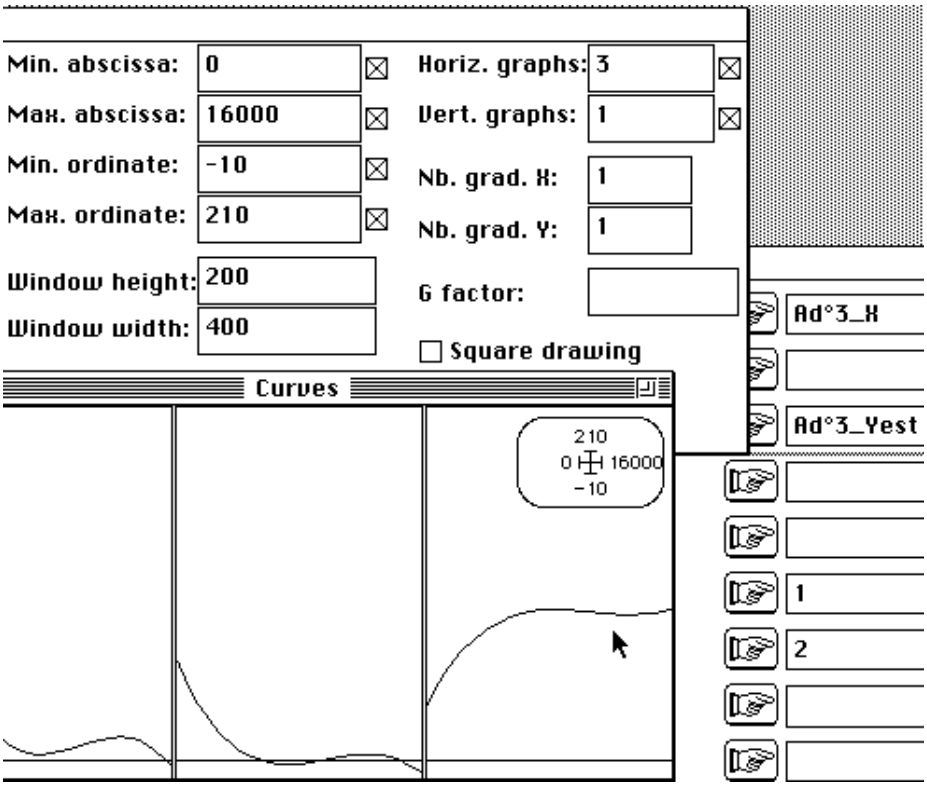
On a :

```
Polynomial model
Y file: Y123
--- Number of rows: 48, columns: 3
X coordinate file: X12
--- Number of rows: 48, columns: 2
--- Selected column: 1
Order of polynomial: 3
-----------------------------------------------
Output file: Ad°3_X
--- Number of rows: 25, column: 1
This file contains the x values used for polynomial estimation
Output file: Ad°3_Yest
--- Number of rows: 25, columns: 3
This file contains the polynomial model for each variable (columns),
-----------------------------------------------
```

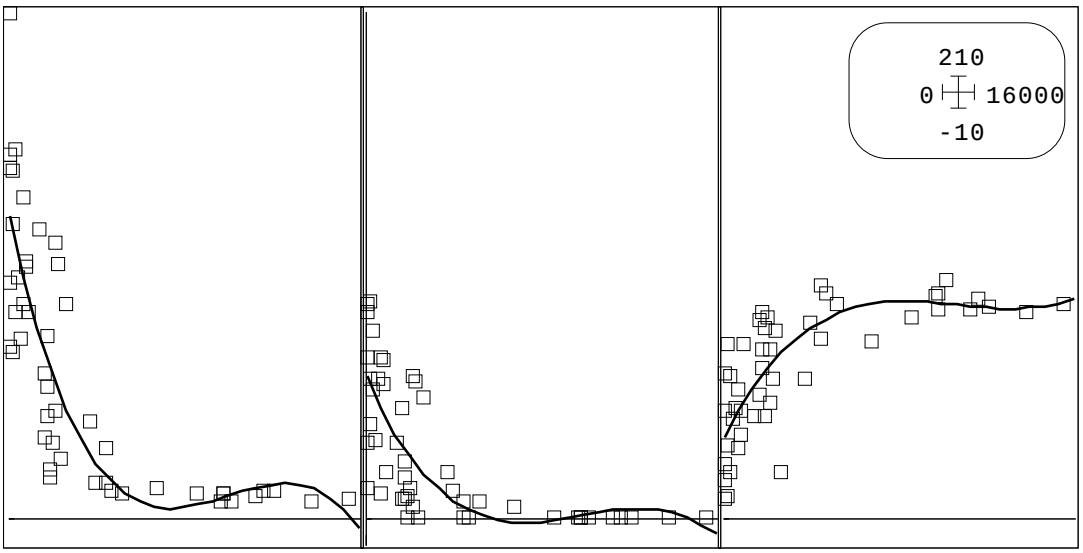
Dans Curves, représenter les données :



Copier la figure, coller dans ClarisDraw™, puis représenter les modèles :



Copier et superposer :



On obtient le même résultat, mais cette procédure est généralisable à tout autre modèle.

2.2 — Changements de variables

Les régressions polynomiales tiennent compte de la non-linéarité mais prennent mal en compte le plateau apparaissant dans la deuxième partie des courbes. On corrige cet élément en faisant un changement de variables. Sélectionner l'explicative et obtenir (Bin->Bin) la variable Log (x) dans le fichier XLog (48-1). Associer XLog et Y dans UnivarReg (option Initialize), puis calculer l'erreur de prédiction.

c*Log[a*x+b]

Input file	X12	48	2
Output file	XLog		
Selection of columns (default = all)	1		
Parameter a (default=1)	1		
Parameter b (default=0)	0		
Parameter c (default=1)	1		

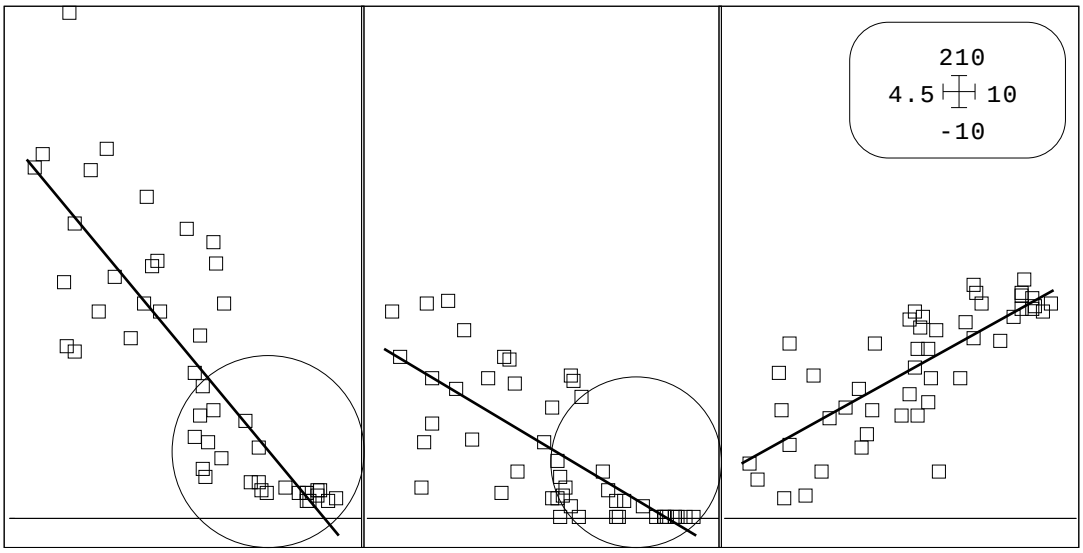
Initialize

Explanatory variable	XLog	48	2
Selected column (default = 1)			
Y file: dependent variables	Y123	48	3
Option: row weight			
Output file name	B		

Polynomial -> Error

Input file	B.uni

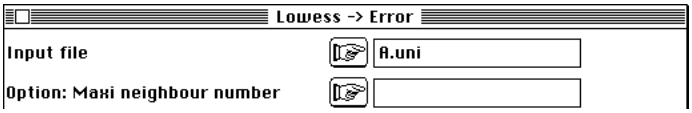
```
Predicted variable n° 1
--------------------------------------------------------------------------
 d°|Error/Vari|    a0    |    a1    |    a2    |    a3    |    a4    |
--------------------------------------------------------------------------
  1| 3.145e-01| 6.212e+01|-4.235e+01|
  2| 3.145e-01| 6.209e+01|-4.234e+01| 3.100e-02|
  3| 3.053e-01| 6.132e+01|-5.431e+01| 2.133e+00| 6.233e+00|
•••
Predicted variable n° 2
--------------------------------------------------------------------------
 d°|Error/Vari|    a0    |    a1    |    a2    |    a3    |    a4    |
--------------------------------------------------------------------------
  1| 4.389e-01| 2.704e+01|-2.114e+01|
  2| 4.380e-01| 2.622e+01|-2.096e+01| 8.243e-01|
•••
Predicted variable n° 3
--------------------------------------------------------------------------
 d°|Error/Vari|    a0    |    a1    |    a2    |    a3    |    a4    |
--------------------------------------------------------------------------
  1| 4.031e-01| 6.210e+01| 1.979e+01|
  2| 4.022e-01| 6.135e+01| 1.996e+01| 7.559e-01|
  3| 3.987e-01| 6.159e+01| 2.370e+01| 9.853e-02|-1.949e+00|
•••
```



Rien n'indique un modèle non linéaire. Calculer les valeurs prédites et représenter données et modèles. Il n'est pas certain que le modèle ainsi obtenu soit optimal. On a souligné par un cercle l'autocorrélation des résidus.

2.3 — Régression locale

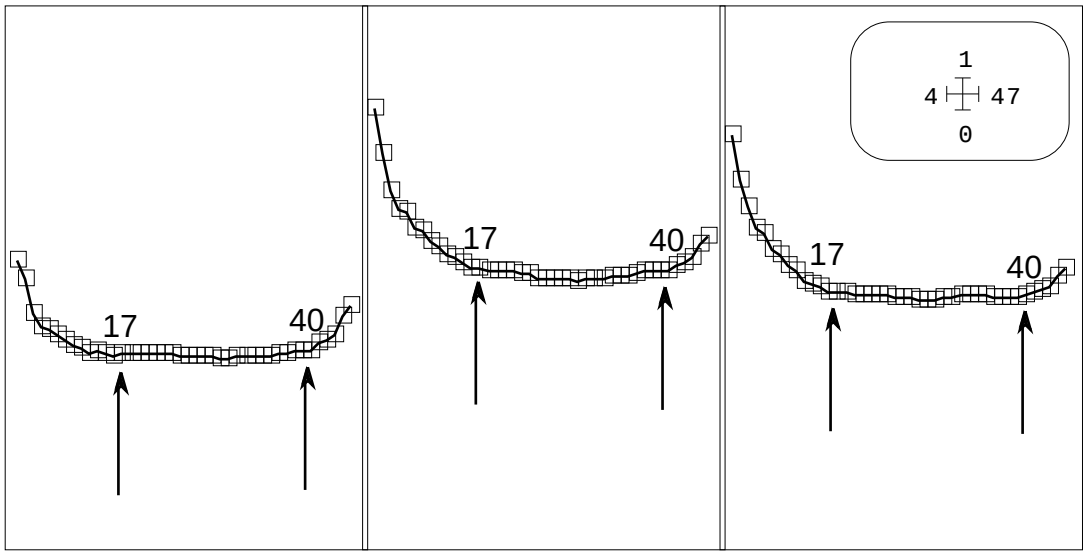
Moins connue est la régression locale ou régression lowess [4]. Au lieu de prendre la totalité des données pour ajuster un polynôme, en particulier une droite, on prend pour chaque point à estimer ses plus proches voisins. On pondère ces voisins par un poids fonction de la distance (fonction tricube), on fait une régression linéaire n'utilisant que ces voisins et on recommence au point suivant. La question est de choisir un nombre de voisins convenable. On étudie l'erreur de prédiction en fonction du nombre dans UniVarReg :



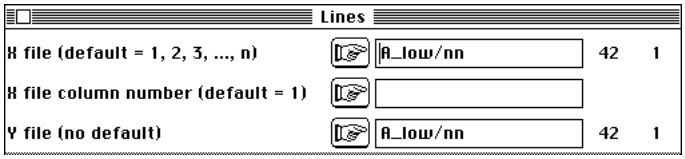
On obtient :

```
lowesserror: total error of 1D-locally weighted regression
X coordinates file: X12
--- Number of rows: 48, columns: 2
--- Selected column: 1
Data file: Y123
--- Number of rows: 48, columns: 3
Maximal neighbour number: 46
Uniform weight
Output file: A_low/err
--- Number of rows: 42, columns: 3
This file contains, for each variable (columns),
and for each number of neighbors (rows) the mean
estimation error for all sampling points
Output file: A_low/nn
--- Number of rows: 42
This file contains the number of neighbours used in each locally weighted
regression
+5.32E-01 +8.18E-01 +7.67E-01
+4.97E-01 +7.34E-01 +6.81E-01
+4.32E-01 +6.61E-01 +6.32E-01
•••
```

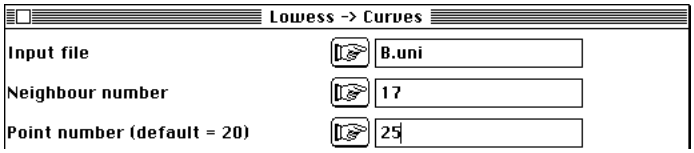
Représenter cette erreur en fonction du nombre de voisins :



avec les paramètres :



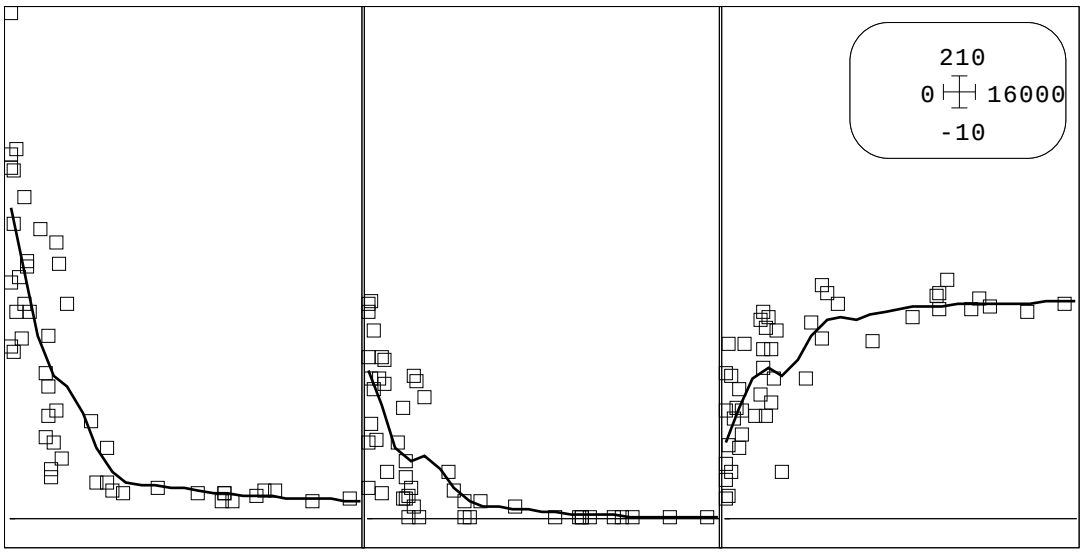
On choisit de minimiser l'erreur en prenant un nombre de voisins le plus petit possible, ici 17. L'erreur est toujours exprimée en pourcentage de la variance. Le similitude des trois courbes est remarquable. On peut alors reconstituer la courbe, comme pour les régressions polynomiales par :



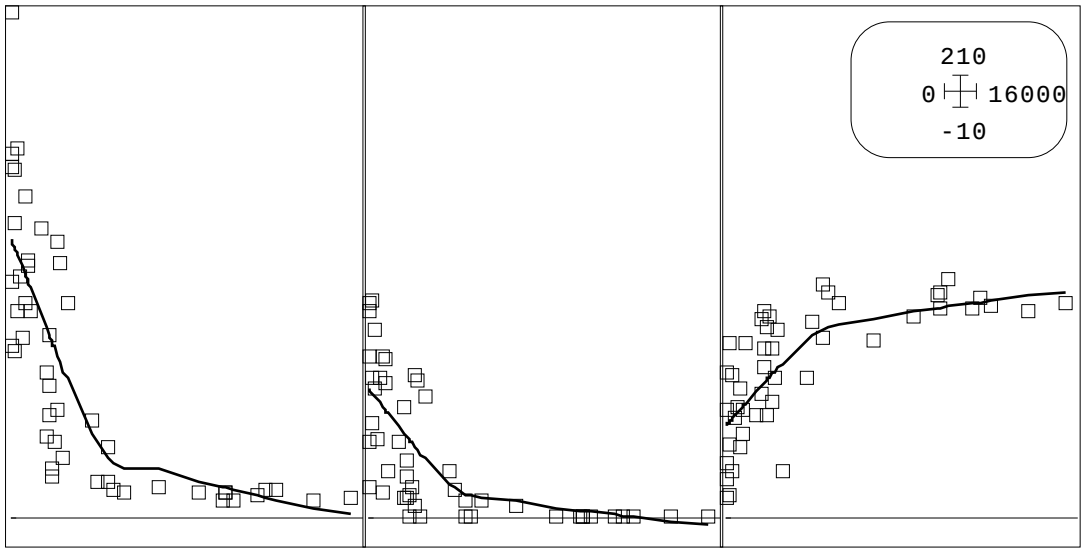
On a :

```
Lowess curves
Y file: Y123
--- Number of rows: 48, columns: 3
X coordinate file: XLog
--- Number of rows: 48, columns: 2
--- Selected column: 1
Number of neighbours used: 17
-----------------------------------------------
Output file: B_low17_X
--- Number of rows: 25, column: 1
This file contains the values (x) used for lowess estimation
Output file: B_low17_Yest
--- Number of rows: 25, columns: 3
This file contains the lowess model for each variable (columns),
-----------------------------------------------
```

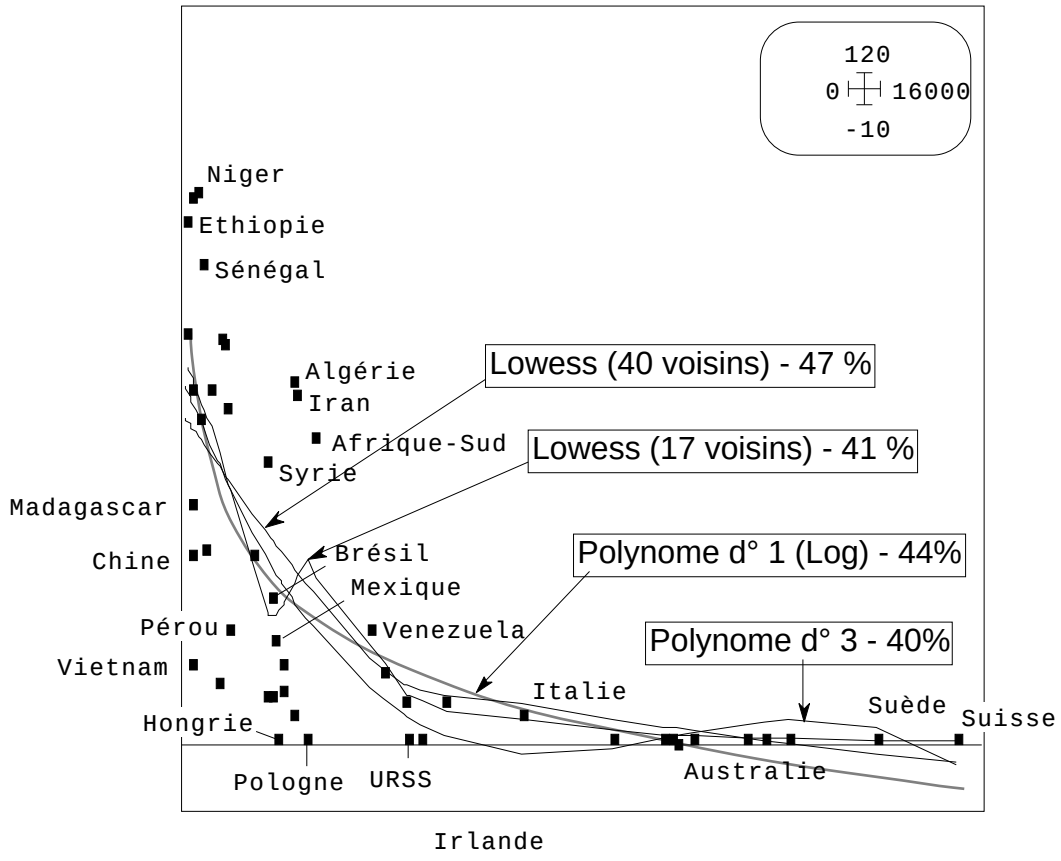
On peut encore superposer données et modèles :



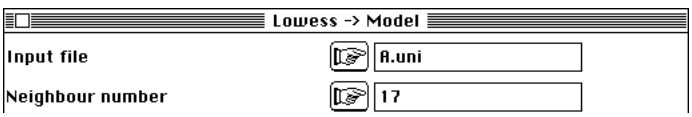
L'attention est attirée sur le décrochement de la courbe sensible dans les trois cas et spécialement sur les deux dernières explicatives. On obtient directement ce graphique dans CurveModels et en utilisant 40 voisins



Ceci montre clairement que la modélisation est un outil d'exploration des données et non un moyen d'obtenir des équations de prédiction aveuglément.



Reprenons la seconde variable. On notera que l'option :



calcule les prévisions aux points de mesure et donne l'erreur :

```
Lowess model
```

```
Y file: Y123
--- Number of rows: 48, columns: 3
X coordinate file: X12
--- Number of rows: 48, columns: 2
--- Selected column: 1
Number of neighbors: 17
Weights = 0.020833
Output file: A_low17
--- Number of rows: 48, columns: 3
This file contains the lowess model for each variable (columns),
Variable  1 | Err/Var   0.283
Variable  2 | Err/Var   0.405
Variable  3 | Err/Var   0.366
```

Dans cette option Lowess -> Model l'erreur (en pourcentage de la variance) tient compte du point lui-même et est comparable à celle du modèle polynomial alors que dans le module Lowess -> Error on n'utilise pas le point de mesure pour effectuer la prévision (comme dans le cas des estimations sur données supplémentaires dans l'option Lowess -> New Data).

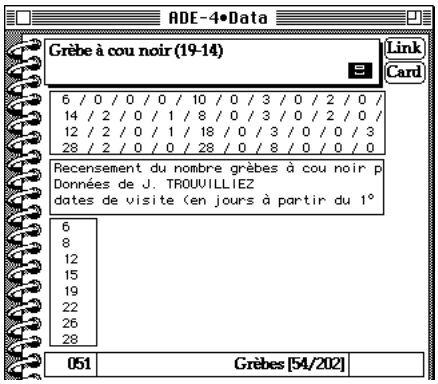
Ce n'est pas dans la minimisation de l'erreur qu'il faut voir un critère absolu de validité des modèles. La part commune souligne la structure des données et l'incompressibilité de la variance résiduelle sur la première moitié de l'intervalle de variation du PIB.

3 — Régression sur vecteurs propres

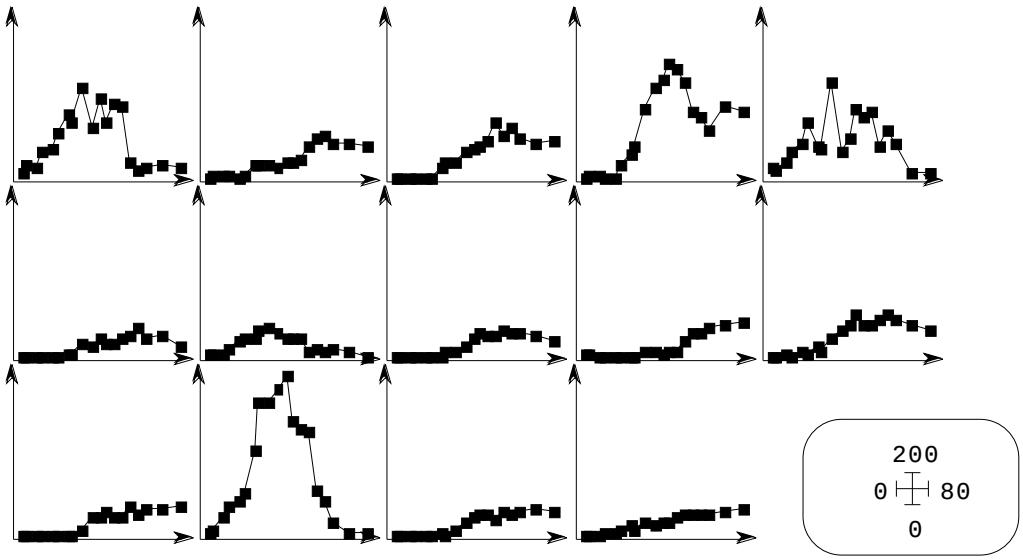
La variable qui est le plus souvent utilisée comme explicative est le temps qui génère les modèles de tendance et de périodicité. ADE-4 n'est pas un logiciel d'analyses de processus temporels, mais nous utiliserons ce type de données pour exprimer les propriétés de la régression sur les vecteurs propres de voisinage qui est une autre manière de générer des modèles non linéaires avec des méthodes linéaires. Les premières données utilisées sont celles de J. Trouvilliez, 1988 [5].

3.1 — Modèles de tendance pour k chroniques

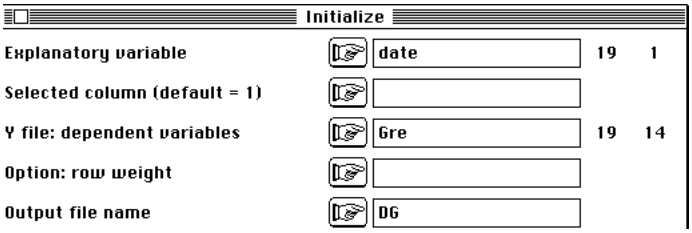
L'auteur a recensé dans quatorze étangs de la plaine du Forez et à dix-neuf reprises dans le temps le nombre de grèbes à cou noir présents sur l'étang (carte Grèbes) :



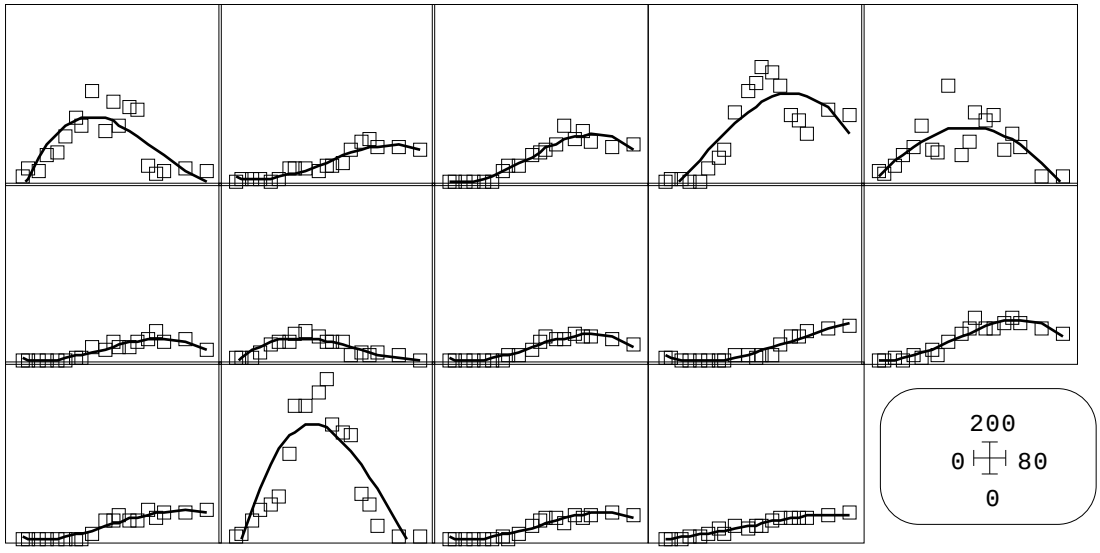
Obtenir les fichiers Gre (19 lignes-dates et 14 colonnes -étangs) et Date (19 lignes-dates et 1 colonne). Représenter les données :



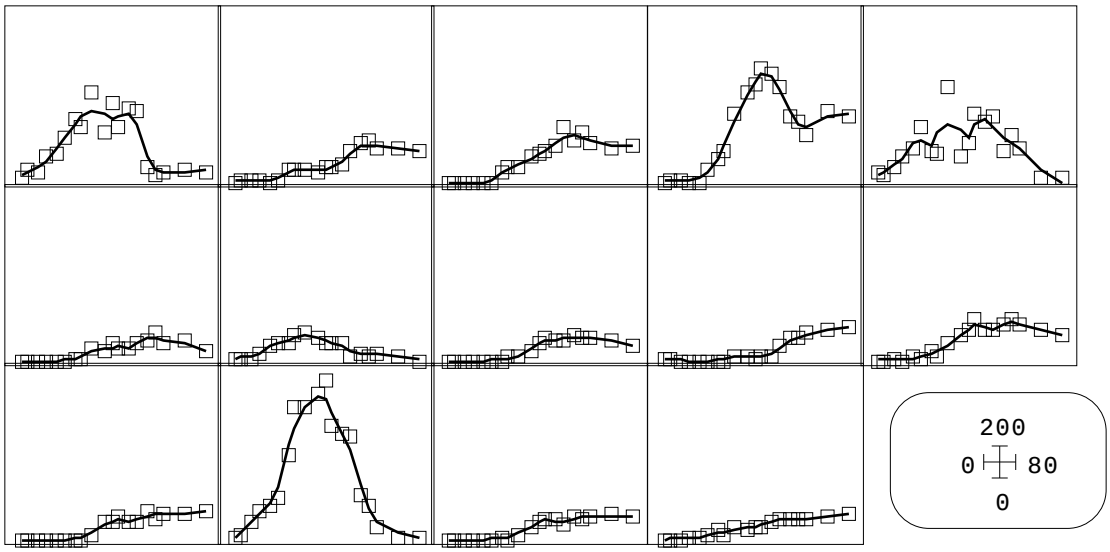
La campagne a commencé le 1er Mars 1985 et a duré 70 jours. On cherche simplement à gommer la variabilité d'échantillonnage en remplaçant les données par des modèles simples. Utiliser d'abord la régression polynomiale.



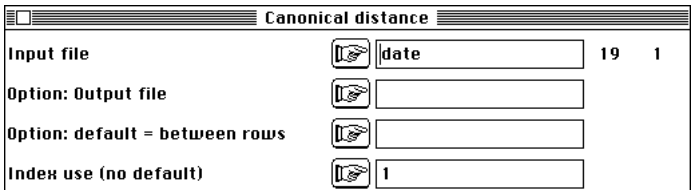
L'option Polynomial -> Error conduit à des polynômes de degré 3 :



L'option lowess -> Error conduit à prendre un petit nombre de voisins (n = 5) pour la régression locale ce qui donne :



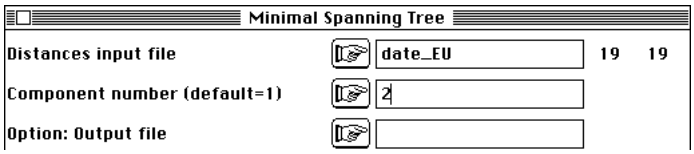
Ni locale, ni polynomiale, la régression linéaire permet d'avoir des modèles curvilignes d'une troisième manière. L'intérêt réside dans le fait qu'il s'agit d'un cas particulier qui permet de prendre en compte des structures très diverses. Procéder de la manière suivante. Utiliser le module Distances :



On a calculé la distance (ici, l'intervalle de temps) qui sépare deux événement :

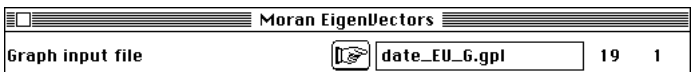
```
Output file: date_EU
It has 19 rows and 19 columns
Distances computed are euclidean
```

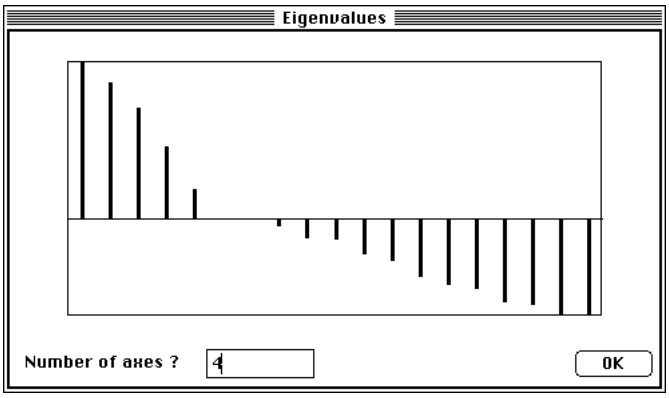
On en extrait un graphe de voisinage dans le même module :




```
Neighboorhing relationship from Minimal Spanning Tree
Input file (distances matrix): date_EU
Rank: 2
Neighbourhing relationship in text file: date_EU_G
It contains graph matrix (LEBART's M) with 19 rows and columns
Neighbourhing weights in binary file: date_EU_G.gpl
It contains  19 rows and 1 column
```

Dans le module NGStat :

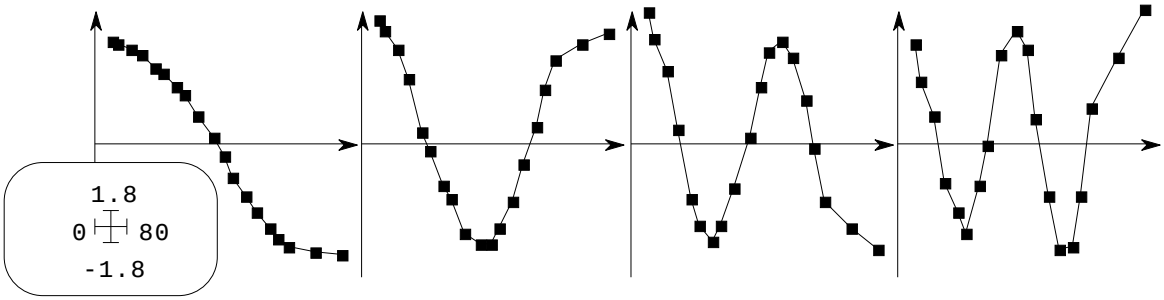




On a diagonalisé l'opérateur de voisinage associé et les vecteurs propres forment un ensemble de codes numériques de référence non corrélées :

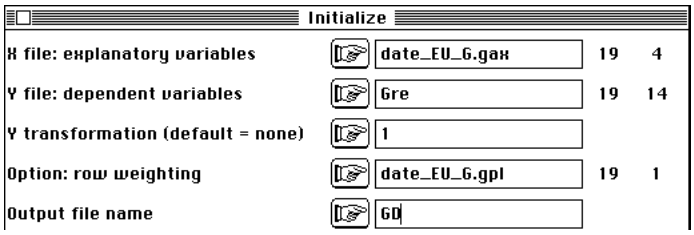
```
Moran operator diagonalization
Access to neighbouring relationship: date_EU_G.gpl
-------------------------------------------
Num  Eigenval.|Num. Eigenval.|Num. Eigenval.|Num. Eigenval.|
001  9.579e-01|002  8.296e-01|003  6.748e-01|004  4.433e-01|
005  1.819e-01|006  1.313e-16|007 -2.980e-08|008 -3.760e-02|
009 -1.137e-01|010 -1.189e-01|011 -2.070e-01|012 -2.500e-01|
013 -3.485e-01|014 -4.006e-01|015 -4.224e-01|016 -5.000e-01|
017 -5.219e-01|018 -5.774e-01|019 -5.894e-01|

File date_EU_G.gvp contains the eigenvalues
--- It has 19 rows and 1 column
File date_EU_G.gax contains the eigenvectors (norm = 1 for weight in file
date_EU_G.gpl)
--- It has 19 rows and 4 columns
```

Représenter ces quatre codes numériques en fonction du temps :



On a choisi de construire une relation de voisinage entre les points de mesure. L'option Table to Distance Matrix a calculé les intervalles de temps séparant deux dates, l'option Minimal Spanning tree a défini une relation de voisinage pour que chaque dates aient au moins deux voisines avant ou après et plus dans les périodes ou le pas d'échantillonnage est réduit, l'option Moran EigenVectors fournit alors quatre fonctions qui ressemblent fortement à quatre polynômes de degré 1, 2, 3 et 4 sans avoir les défauts de ne pas être bornés et d'être corrélés, ce qui serait intervenu si on avait pris x, x^2, x^3 et x^4 (x étant la date). La modélisation de chaque chronique à partir de ces pseudo-polynômes orthogonaux utilise le module OrthoVar :

Initialiser l'association explicatives (Date_EU$Gax) et expliquées (Gre) par l'option Initialize :



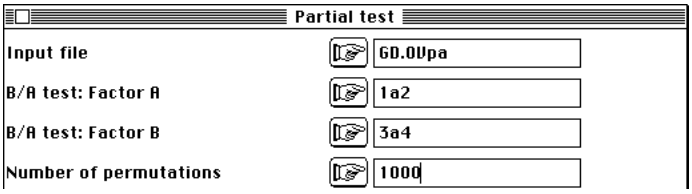

```
New TEXT file GD.OVpa contains the parameters:
----> Explanatory variables: date_EU_G.gax [19][4]
----> Dependent variable file: Gre [19][14]
----> Transformation used: 1
      0 = None 1 = D-centring, 2 = D-standardization, 3 = D-normalization
----> Row weight file: date_EU_G.gpl
-----------------------------------------------------
File GD.OVcs contains the cosinus square between explanatory and dependent
variables:
----> 4 rows (explanatory variables)
----> 14 columns (dependent variables)
*-----------------------------------*
| N° | Variance | Explained|  Ratio   |
|   1| 9.940e+02| 7.884e+02| 7.932e-01|
|   2| 2.401e+02| 2.178e+02| 9.071e-01|
|   3| 4.317e+02| 4.073e+02| 9.435e-01|
|   4| 2.095e+03| 2.024e+03| 9.664e-01|
|   5| 6.915e+02| 3.519e+02| 5.090e-01|
|   6| 1.016e+02| 8.492e+01| 8.362e-01|
|   7| 1.000e+02| 9.036e+01| 9.036e-01|
|   8| 1.309e+02| 1.272e+02| 9.712e-01|
|   9| 1.526e+02| 1.484e+02| 9.724e-01|
|  10| 3.054e+02| 2.860e+02| 9.366e-01|
|  11| 1.583e+02| 1.488e+02| 9.403e-01|
|  12| 3.737e+03| 3.518e+03| 9.414e-01|
|  13| 1.249e+02| 1.209e+02| 9.677e-01|
|  14| 8.986e+01| 8.679e+01| 9.659e-01|
*-----------------------------------*
```

La colonne de droite donne le pourcentage de variance expliquée par la régression sur les quatre variables (minimum 51% pour la courbe 5, maximum 97% pour la 4). On peut s'interroger sur le rôle de chacune des explicatives dans la modélisation de chacune des chroniques. Les explicatives étant non corrélées, il n'y a aucune ambiguïté, car la variance expliquée s'additionne purement et simplement et on peut décortiquer la constitution du pourcentage de variance expliquée. Le module Variable test permet en outre de fournir un seuil de signification par un test de permutation.


```
----> Explanatory variables: date_EU_G.gax
----> Dependent variable file: Gre
----> Transformation used: 1
      0 = None 1 = D-centring, 2 = D-standardization, 3 = D-normalization
----> Row weight file: date_EU_G.gpl
----> Number of random permutations: 1000
-----------------------------------------------------
Dependent variable number: 1
----------------------------------
|VarX|r2 observ.|X>Xobs| Frequency|
----------------------------------
|   1| 4.966e-03|   778| 7.780e-01|
|   2| 7.868e-01|     0| 0.000e+00|
|   3| 1.448e-03|   862| 8.620e-01|
|   4| 3.807e-05|   980| 9.800e-01|
----------------------------------
```

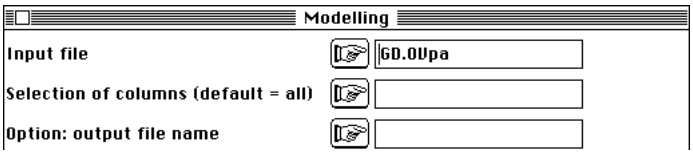
```
Dependent variable number: 2
---------------------------------
|VarX|r2 observ.|X>Xobs| Frequency|
---------------------------------
|   1| 8.709e-01|     0| 0.000e+00|
|   2| 1.519e-02|   592| 5.920e-01|
|   3| 4.983e-03|   768| 7.680e-01|
|   4| 1.607e-02|   590| 5.900e-01|
---------------------------------
Dependent variable number: 3
---------------------------------
|VarX|r2 observ.|X>Xobs| Frequency|
---------------------------------
|   1| 8.902e-01|     0| 0.000e+00|
|   2| 1.872e-02|   574| 5.740e-01|
|   3| 2.683e-02|   474| 4.740e-01|
|   4| 7.732e-03|   701| 7.010e-01|
---------------------------------
•••
```

On observe que soit la première soit la seconde des explicatives jouent un rôle essentiel. Ces deux premières variables sont retenues pour modéliser simultanément les 14 chroniques. On peut se demander si les corrections qu'apporteraient aux modèles les deux dernières sont utiles. L'option Partial test permet de répondre à cette question par un test de permutations radicalement différent du précédent :



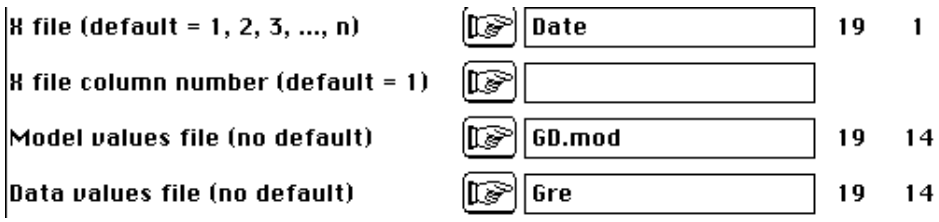

```
----> Explanatory variables: date_EU_G.gax
----> Dependent variable file: Gre
----> Transformation used: 1
      0 = None 1 = D-centring, 2 = D-standardization, 3 = D-normalization
----> Row weight file: date_EU_G.gpl
----> Number of random permutations: 1000
----> Effect A: selection of explanatory variables: 1a2
----> Effect B: selection of explanatory variables: 3a4
-----------------------------------------------------
-----------------------------------------------------------
|VarY|r2 observ.| mean sim.| normal I |X>Xobs| Frequency|
-----------------------------------------------------------
|   1| 7.137e-03| 1.117e-01|-1.039e+00|   945| 9.450e-01|
|   2| 1.848e-01| 1.091e-01| 7.480e-01|   178| 1.780e-01|
|   3| 3.796e-01| 1.142e-01| 2.637e+00|    23| 2.300e-02|
|   4| 7.884e-01| 1.072e-01| 6.946e+00|     0| 0.000e+00|
|   5| 1.400e-01| 1.072e-01| 3.450e-01|   282| 2.820e-01|
|   6| 1.036e-01| 1.084e-01|-5.013e-02|   410| 4.100e-01|
|   7| 4.974e-01| 1.116e-01| 4.058e+00|     3| 3.000e-03|
|   8| 7.436e-01| 1.076e-01| 6.839e+00|     0| 0.000e+00|
|   9| 5.382e-01| 1.108e-01| 4.587e+00|     0| 0.000e+00|
|  10| 3.931e-01| 1.099e-01| 2.840e+00|    23| 2.300e-02|
|  11| 3.740e-01| 1.092e-01| 2.602e+00|    27| 2.700e-02|
|  12| 3.358e-01| 1.083e-01| 2.409e+00|    25| 2.500e-02|
|  13| 5.674e-01| 1.102e-01| 4.670e+00|     0| 0.000e+00|
|  14| 4.022e-02| 1.063e-01|-7.114e-01|   717| 7.170e-01|
-----------------------------------------------------------
```

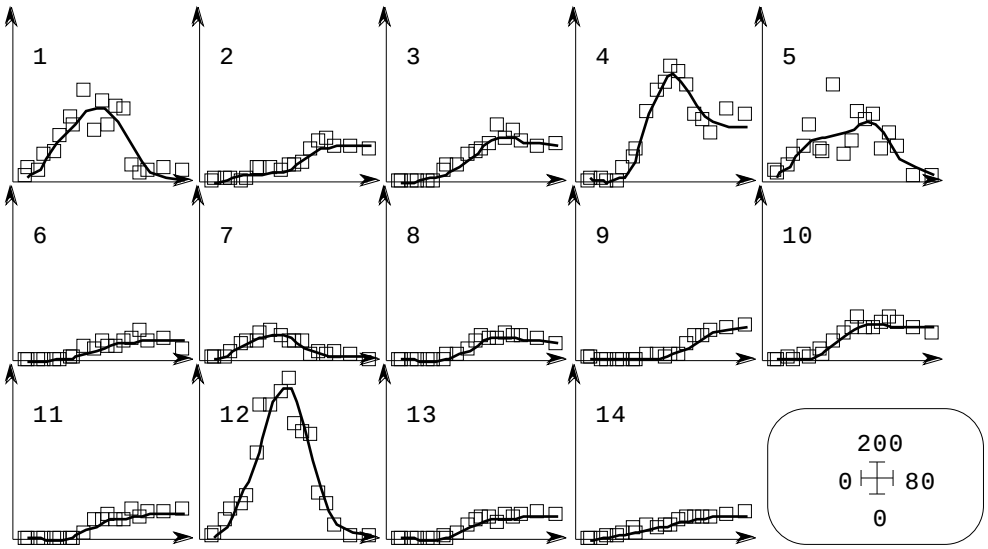
On permute ici les résidus de la prédiction par les variables 1 et 2 pour tester strictement l'effet des variables 3 et 4 sur le reste. L'effet n'est pas négligeable dans plusieurs des cas et on décide de conserver les quatre variables prédictrices.




```
----> Explanatory variables: date_EU_G.gax
----> Dependent variable file: Gre
----> Transformation used: 1
      0 = None 1 = D-centring, 2 = D-standardization, 3 = D-normalization
----> Row weight file: date_EU_G.gpl
----> Selection of explanatory variables: 1a4
-------------------------------------------------------
File GD.mod has 19 rows and 14 columns
It contains the linear models resulting
from the separate multiple linear regression of each dependent variable
upon the set of explanatory variables
File :GD.mod
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1| 1.430e-01| 8.518e+01|
|   2|-2.499e+00| 4.114e+01|
|   3|-2.511e+00| 5.147e+01|
|   4|-4.414e+00| 1.233e+02|
|   5| 5.699e+00| 6.871e+01|
|   6|-4.611e-01| 2.174e+01|
|   7| 2.684e-01| 2.942e+01|
|   8|-3.961e-01| 2.753e+01|
|   9| 7.227e-02| 3.804e+01|
|  10| 3.514e-01| 4.140e+01|
|  11|-2.018e+00| 3.076e+01|
|  12| 2.405e+00| 1.750e+02|
|  13|-1.822e+00| 2.705e+01|
|  14|-1.243e+00| 2.653e+01|
|----|----------|----------|
File GD.res has 19 rows and 14 columns
It contains (data - model) matrix
File :GD.res
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-2.686e+01| 2.827e+01|
|   2|-7.233e+00| 7.589e+00|
|   3|-7.139e+00| 1.092e+01|
|   4|-1.492e+01| 1.859e+01|
|   5|-2.765e+01| 5.542e+01|
|   6|-1.066e+01| 9.262e+00|
|   7|-6.222e+00| 5.899e+00|
|   8|-4.646e+00| 3.789e+00|
|   9|-4.840e+00| 5.382e+00|
|  10|-8.842e+00| 8.305e+00|
|  11|-5.534e+00| 7.133e+00|
|  12|-2.595e+01| 3.561e+01|
|  13|-4.567e+00| 3.328e+00|
|  14|-3.165e+00| 3.630e+00|
|----|----------|----------|
```

On utilise CurveModels pour superposer données et modèles :

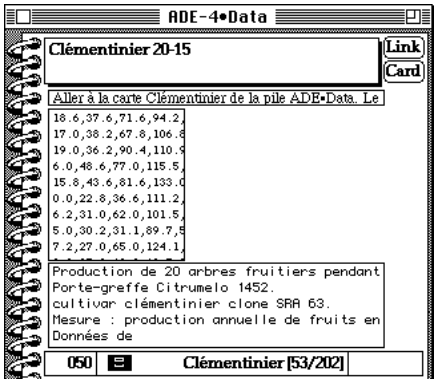




Les modèles résument parfaitement l'évolution des chroniques. Nous retiendrons de cela qu'il convient de ne pas confondre logique de régression linéaire (au sens algèbre linéaire) et courbe de réponses linéaire (au sens de en forme de droite). Il semble justifier de préférer régression sur vecteurs propres indépendants à régression polynomiale car le principe de telles pratiques s'étend à nombre de situations.

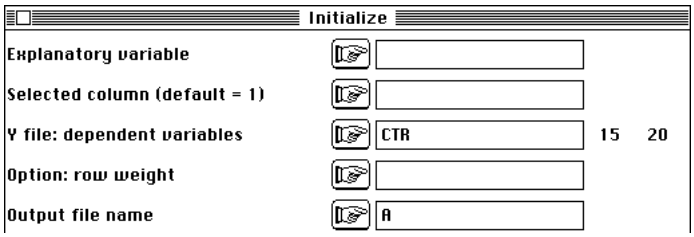
3.2 — Modèles d'alternance de k chroniques

La méthode s'étend à un problème beaucoup plus complexe posé par D. Tisne-Agostini [6] pour l'analyse d'un vaste ensemble de données d'un essai porte-greffe de la Station de Recherches Agronomiques de San Giuliano [7] . La méthodologie en œuvre est voisine de celle de Méot & Coll. [8], à une variante de centrage près. Les données traitées sont dans la carte Clémentinier de la pile ADE-4•Data :



Le tableau comporte 20 lignes et 15 colonnes et donne la production annuelle de fruits en kg de 20 arbres fruitiers pendant 15 années. Faire avec le champ du haut un fichier C.edit , le passer en binaire sous le nom C (20-15) et le transposer sous le nom CTR (15-20). On obtient 20 courbes (colonnes) sur 15 points (lignes). Obtenir une représentation des données brutes (Figure 1).

Ces données contiennent deux éléments de modélisation. D'une part la production augmente avec l'age de l'arbre, d'autre part le phénomène d'alternance est sensible, surtout dans la seconde partie de la courbe. Enlevons la tendance. Dans UniVarReg, on peut ne pas donner de fichier X :



Implicitement la variable explicative est la numérotation naturelle :

```
-------------------------------------------------------
New TEXT file A.uni contains the parameters:
----> Explanatory variables: A.12...n [15][1]
----> Selected variable: 1
----> Dependent variable file: CTR [15][20]
----> Row weight file: Uniform_weight
-------------------------------------------------------
```

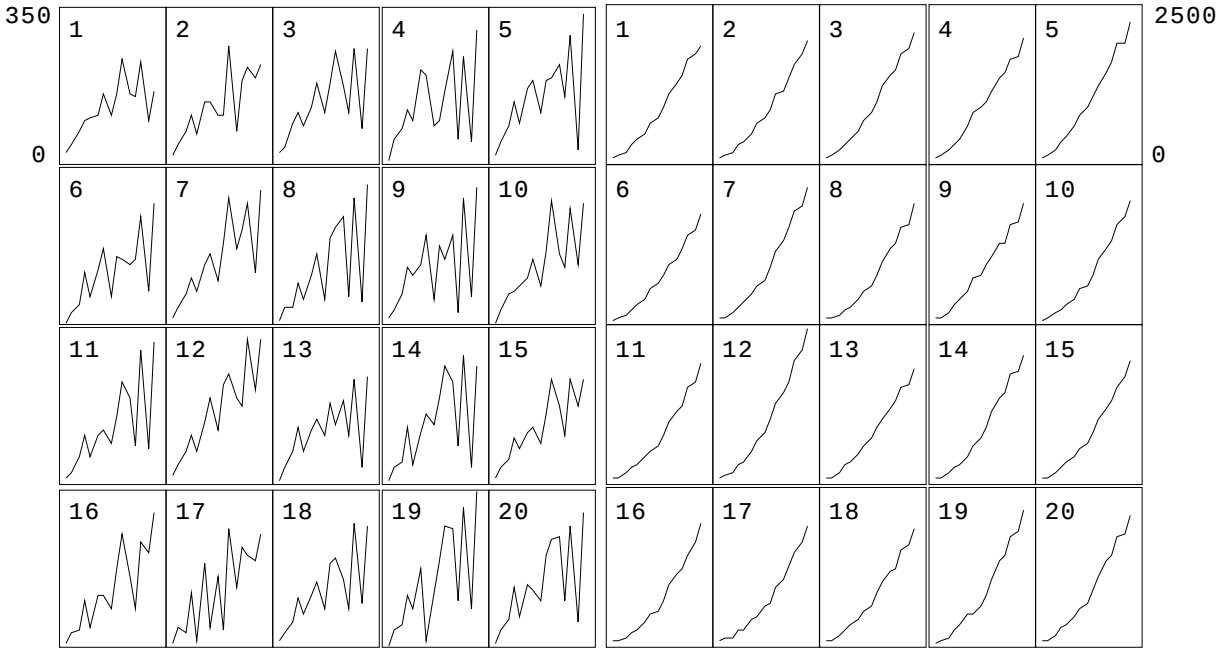
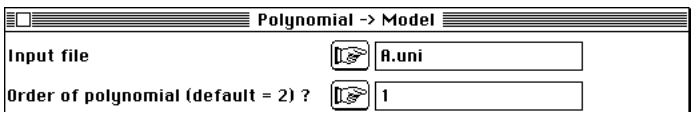


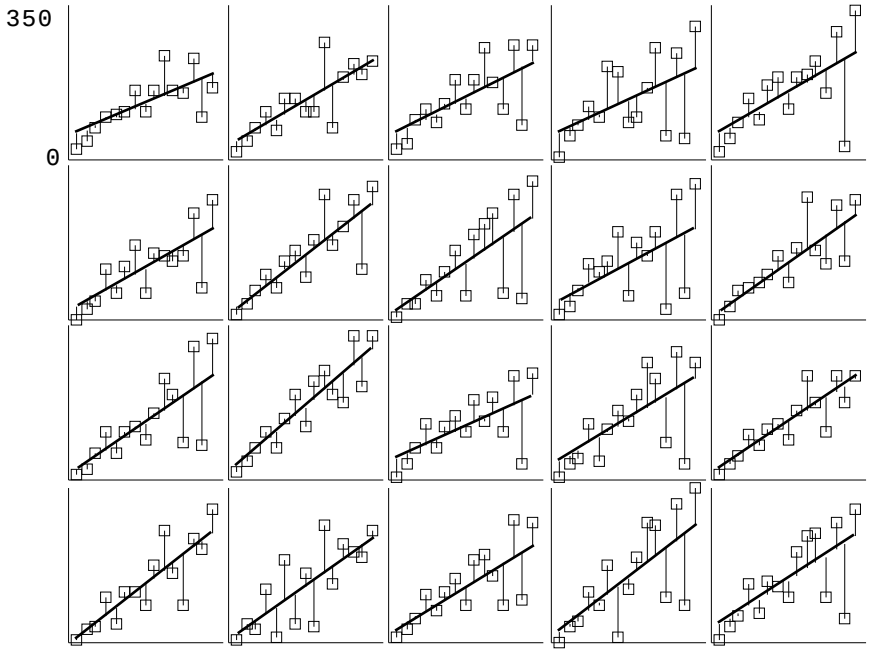
Figure 1 : Production en kg de fruits de 20 arbres fruitiers pendant 15 ans. Représentation simple à gauche et cumulée à droite. On voit la croissance de la production et l'alternance des années de forte et faible production.

On utilise simplement un polynôme de degré 1 dans Polynomial -> model :

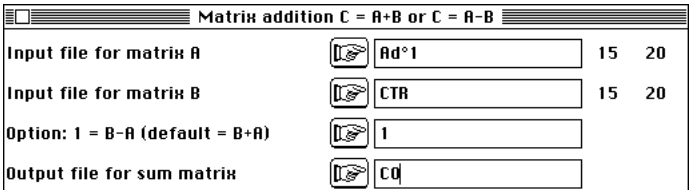



```
Polynomial model
Y file: CTR
--- Number of rows: 15, columns: 20
X coordinate file: A.12...n
--- Number of rows: 15, columns: 1
--- Selected column: 1
Order of polynomial: 1
Output file: Ad°1
--- Number of rows: 15, columns: 20
This file contains the polynomial model for each variable (columns),
```

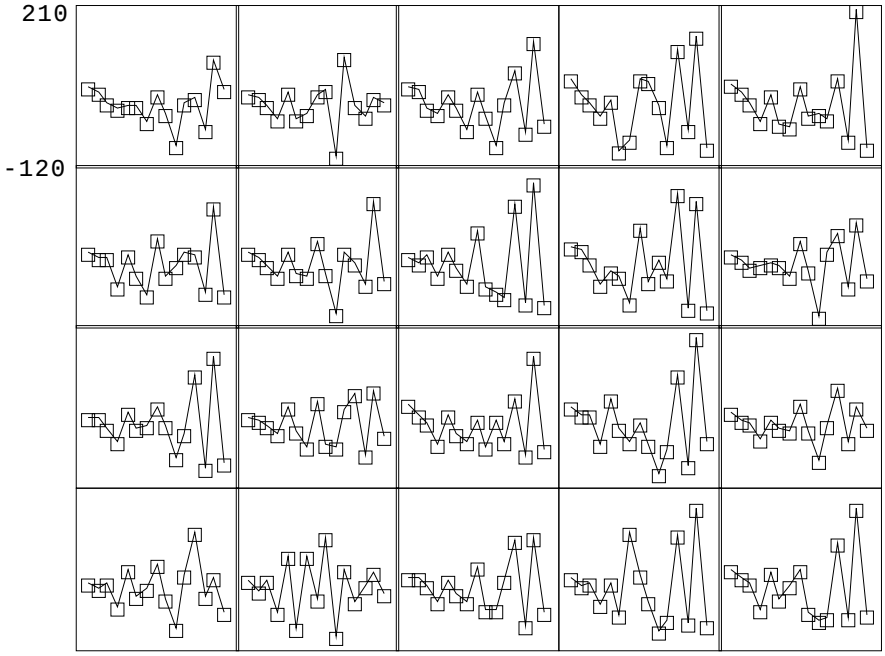
L’opération est représentée dans CurveModels :



La différence entre données (CTR) et modèle (Ad°1) se calcule dans MatAlg :

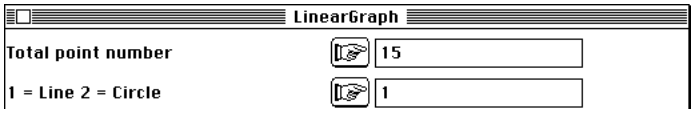


Représenter le nouveau tableau :



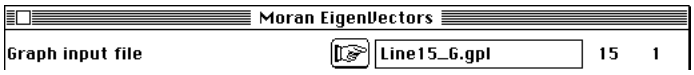
Nous sommes en présence d'une relation de voisinage linéaire simplement définie par la relation : deux points sont voisins si ils correspondent à deux années

consécutives. Implanter cette relation de voisinage par l'option LinearGraph de NGUtil :



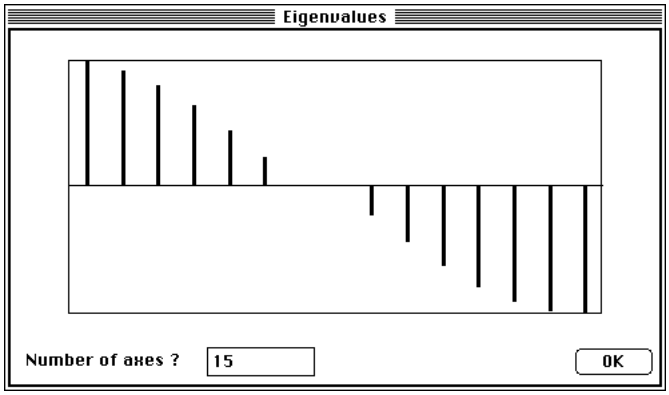

```
Geometric neighbouring relationship
15 points on a line
Neighbourhing relationship in text file: Line15_G
It contains graph matrix (LEBART's M) with 15 rows and columns
Neighbourhing weights in binary file: Line15_G.gpl
It contains  15 rows and 1 column
```

Comme dans le cas précédent, utiliser NGStat :

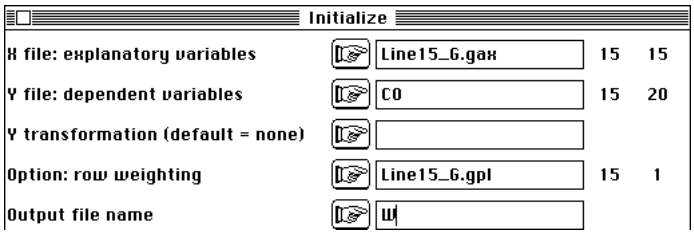



```
Moran operator diagonalization
Access to neighbouring relationship: Line15_G.gpl
-----------------------------------------
Num  Eigenval.|Num. Eigenval.|Num. Eigenval.|Num. Eigenval.|
001  9.749e-01|002  9.010e-01|003  7.818e-01|004  6.235e-01|
005  4.339e-01|006  2.225e-01|007  6.146e-17|008 -8.941e-08|
009 -2.225e-01|010 -4.339e-01|011 -6.235e-01|012 -7.818e-01|
013 -9.010e-01|014 -9.749e-01|015 -1.000e+00|

File Line15_G.gvp contains the eigenvalues
--- It has 15 rows and 1 column
File Line15_G.gax contains the eigenvectors (norm = 1 for weight in file
Line15_G.gpl)
--- It has 15 rows and 15 columns
```

On a gardé cette fois-ci tous les axes :



Le programme crée 15 codes numériques (vecteurs propres de l'opérateur de lissage, cf. théorie dans Thioulouse & Coll [9]). Les premiers sont lisses et les derniers de plus en plus auto corrélés négativement. Utiliser le fichier L15_G.gax pour visualiser ces courbes et leur propriété fondamentale (Figure 3). Pour étudier la prédiction des données, coupler le tableau à expliquer (C0) et le tableau des explicatives (L15_G.gax) dans OrthoVar :



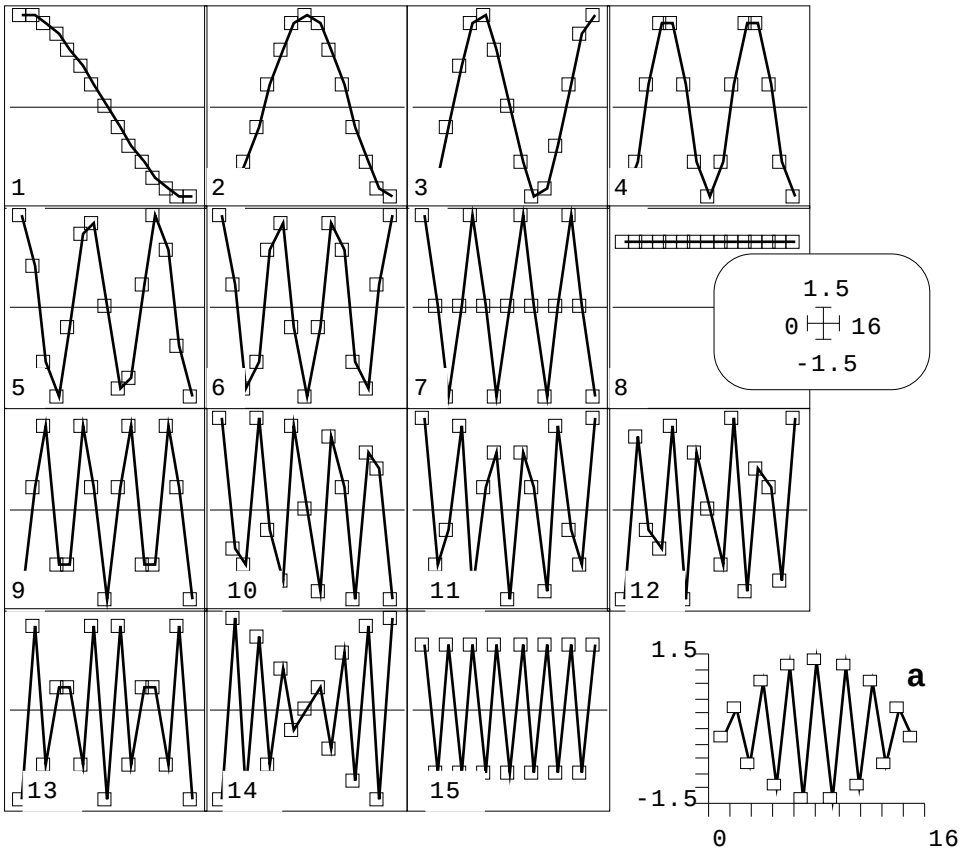


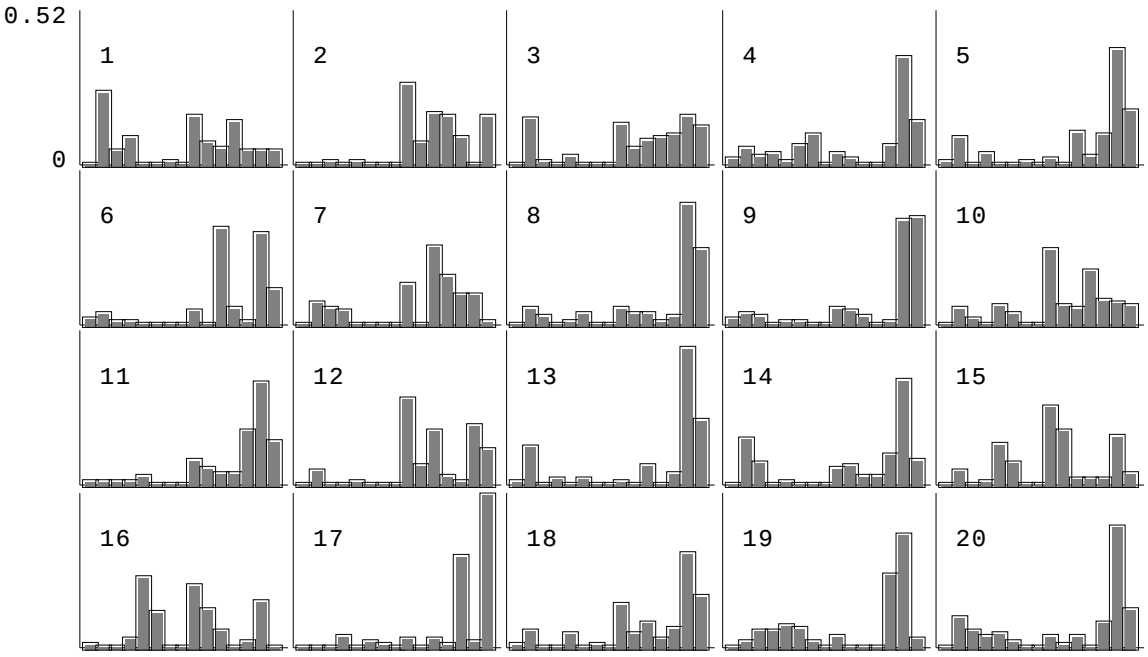
Figure 3 : Vecteurs propres de l'opérateur de lissage pour la relation de voisinage linéaire à 15 points. Ces codes de référence par régression multiple sont aptes à modéliser la croissance pour les premiers et l'alternance pour les derniers. ***a*** *- Pour comparaison, le dernier vecteur propre de l'opérateur de voisinage utilisé dans la version 3.7 d'ADE (Méot & Coll. op. cit.)*

On a :

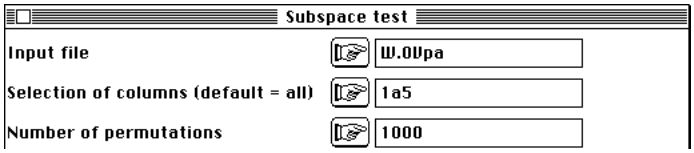
```
New TEXT file W.OVpa contains the parameters:
----> Explanatory variables: Line15_G.gax [15][15]
----> Dependent variable file: C0 [15][20]
----> Transformation used: 0
      0 = None 1 = D-centring, 2 = D-standardization, 3 = D-normalization
----> Row weight file: Line15_G.gpl
-------------------------------------------------------
File W.OVcs contains the cosinus square between explanatory and dependent
variables:
----> 15 rows (explanatory variables)
----> 20 columns (dependent variables)
*-------------------------------------*
| N° | Variance | Explained|  Ratio   |
|   1| 1.683e+03| 1.683e+03| 1.000e+00|
|   2| 1.931e+03| 1.931e+03| 1.000e+00|
|   3| 3.005e+03| 3.005e+03| 1.000e+00|
|   4| 5.262e+03| 5.262e+03| 1.000e+00|
•••
|  18| 2.740e+03| 2.740e+03| 1.000e+00|
|  19| 6.064e+03| 6.064e+03| 1.000e+00|
|  20| 4.385e+03| 4.385e+03| 1.000e+00|
*-------------------------------------*
```

La base orthonormale de vecteurs propres étant complète, le pourcentage de variance expliquée est mathématiquement égale à 1 (100%). L'objectif n'est évidemment pas de reconstituer intégralement les données mais de les modéliser.

On examine d'abord quelles sont les variables en cause, la variance se décomposant en parties additives associées à chaque variable prédictrice. Utiliser Curves sur le fichier W.OVcs pour visualiser directement cette décomposition :



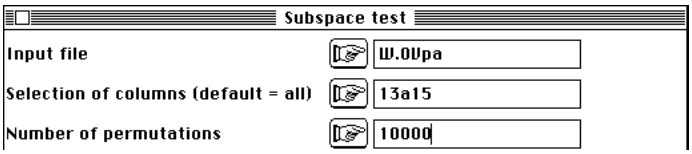
On observe la présence de la plus grande partie de la prédiction dans les derniers vecteurs propres (alternance). Un test de l'effet des cinq premiers vecteurs propres est exécuté par :



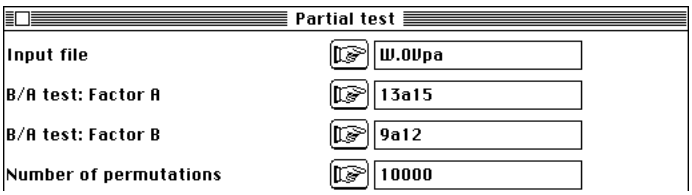
VarY	r2 observ.	mean sim.	normal I	X>Xobs	Frequency
1	3.940e-01	3.580e-01	2.195e-01	397	3.970e-01
2	3.672e-02	3.588e-01	-2.114e+00	996	9.960e-01
3	2.219e-01	3.562e-01	-8.114e-01	766	7.660e-01
4	1.725e-01	3.498e-01	-1.081e+00	852	8.520e-01
5	1.500e-01	3.539e-01	-1.369e+00	916	9.160e-01
6	1.007e-01	3.540e-01	-1.575e+00	954	9.540e-01
7	1.936e-01	3.575e-01	-1.072e+00	830	8.300e-01
8	1.186e-01	3.597e-01	-1.398e+00	945	9.450e-01
9	1.105e-01	3.535e-01	-1.443e+00	949	9.490e-01
10	1.619e-01	3.554e-01	-1.184e+00	879	8.790e-01
11	8.379e-02	3.553e-01	-1.643e+00	971	9.710e-01
12	7.996e-02	3.576e-01	-1.585e+00	969	9.690e-01
13	1.693e-01	3.541e-01	-1.183e+00	874	8.740e-01
14	2.567e-01	3.593e-01	-6.268e-01	698	6.980e-01
15	2.084e-01	3.613e-01	-9.328e-01	803	8.030e-01
16	2.845e-01	3.590e-01	-4.559e-01	642	6.420e-01
17	4.477e-02	3.597e-01	-1.804e+00	993	9.930e-01
18	1.137e-01	3.557e-01	-1.419e+00	939	9.390e-01
19	2.217e-01	3.575e-01	-7.865e-01	763	7.630e-01
20	2.542e-01	3.600e-01	-6.296e-01	696	6.960e-01

On trouve dans ce tableau le r2 observé de la régression sur les cinq premiers codes, puis pour le nombre de permutations demandées la moyenne des r2 observés, la valeur normalisée (valeur observée - valeur moyenne simulée divisée par écart-type simulé), la

fréquence dépassement de l'observation dans l'ensemble des permutations. Aucune variable ne propose un quelconque niveau de signification. On recommence avec les trois dernières :

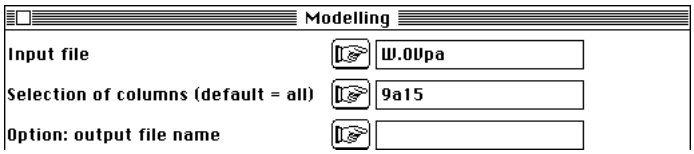


VarY	r2 observ.	mean sim.	normal I	X>Xobs	Frequencie
1	1.484e-01	2.181e-01	-4.901e-01	6230	6.230e-01
2	2.641e-01	2.168e-01	3.521e-01	3270	3.270e-01
3	4.007e-01	2.183e-01	1.274e+00	1178	1.178e-01
4	5.834e-01	2.157e-01	2.504e+00	202	2.020e-02
5	6.714e-01	2.164e-01	3.509e+00	9	9.000e-04
6	4.467e-01	2.183e-01	1.600e+00	770	7.700e-02
7	2.206e-01	2.191e-01	1.147e-02	4429	4.429e-01
8	6.929e-01	2.179e-01	3.318e+00	20	2.000e-03
9	7.326e-01	2.168e-01	3.532e+00	17	1.700e-03
10	2.301e-01	2.189e-01	7.918e-02	4076	4.076e-01
11	6.832e-01	2.183e-01	3.256e+00	33	3.300e-03
12	3.397e-01	2.188e-01	8.033e-01	1995	1.995e-01
13	7.283e-01	2.162e-01	3.787e+00	5	5.000e-04
14	5.487e-01	2.183e-01	2.352e+00	235	2.350e-02
15	2.218e-01	2.183e-01	2.485e-02	4330	4.330e-01
16	1.896e-01	2.178e-01	-2.036e-01	5152	5.152e-01
17	8.453e-01	2.162e-01	4.270e+00	0	0.000e+00
18	5.627e-01	2.185e-01	2.344e+00	236	2.360e-02
19	6.549e-01	2.169e-01	3.009e+00	56	5.600e-03
20	6.204e-01	2.172e-01	2.868e+00	89	8.900e-03



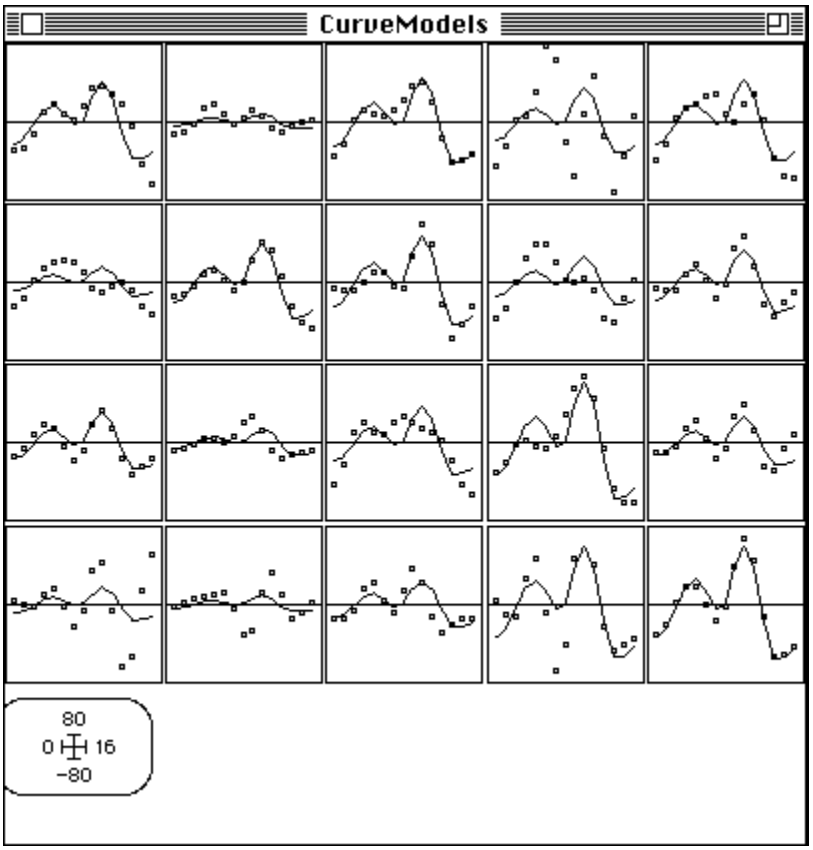
VarY	r2 observ.	mean sim.	normal I	X>Xobs	Frequencie
1	5.157e-01	2.824e-01	1.525e+00	822	8.220e-02
2	9.423e-01	2.851e-01	4.195e+00	0	0.000e+00
3	6.227e-01	2.826e-01	2.254e+00	241	2.410e-02
4	1.712e-01	2.804e-01	-6.875e-01	7138	7.138e-01
5	5.056e-01	2.811e-01	1.410e+00	1006	1.006e-01
6	8.077e-01	2.841e-01	3.256e+00	14	1.400e-03
7	7.441e-01	2.842e-01	3.011e+00	31	3.100e-03
8	4.780e-01	2.818e-01	1.407e+00	948	9.480e-02
9	5.412e-01	2.818e-01	1.610e+00	765	7.650e-02
10	7.319e-01	2.837e-01	3.260e+00	13	1.300e-03
11	7.082e-01	2.809e-01	2.967e+00	35	3.500e-03
12	8.734e-01	2.846e-01	3.669e+00	0	0.000e+00
13	2.995e-01	2.812e-01	1.135e-01	4074	4.074e-01
14	4.256e-01	2.821e-01	9.796e-01	1712	1.712e-01
15	6.361e-01	2.847e-01	2.470e+00	109	1.090e-02
16	5.015e-01	2.856e-01	1.440e+00	912	9.120e-02
17	5.009e-01	2.845e-01	1.487e+00	846	8.460e-02
18	7.014e-01	2.835e-01	2.659e+00	90	9.000e-03
19	1.145e-01	2.800e-01	-1.079e+00	8550	8.550e-01
20	2.790e-01	2.817e-01	-1.776e-02	4636	4.636e-01

Le test partiel indique clairement que sur les 20 chroniques 14 d'entre elles (en italique ci-dessus) ont des relations très significatives avec l'ensemble des vecteurs propres d'autocorrélation négative de rang 9 à 15. Le sous-espace des 7 derniers vecteurs est utilisé pour débarrasser les données de l'autocorrélation négative :

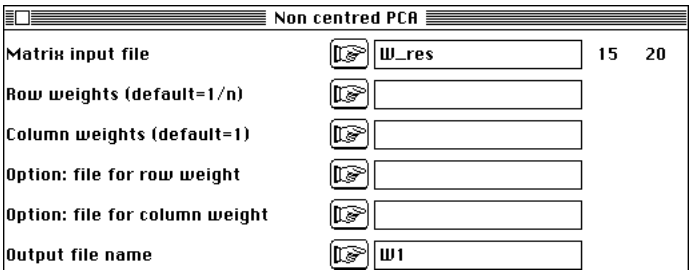


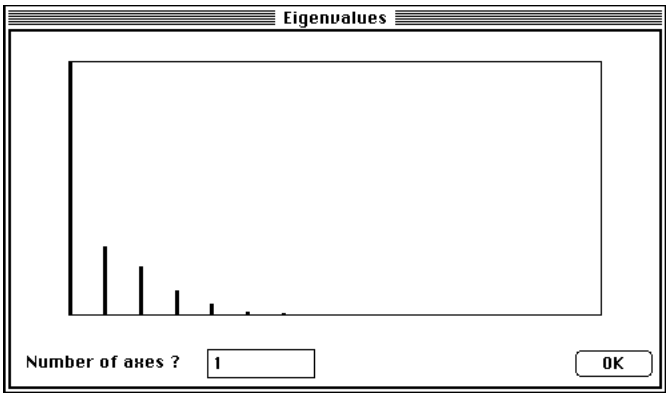

```
----> Explanatory variables: Line15_G.gax
----> Dependent variable file: C0
----> Transformation used: 0
      0 = None 1 = D-centring, 2 = D-standardization, 3 = D-normalization
----> Row weight file: Line15_G.gpl
----> Selection of explanatory variables: 9a15
-------------------------------------------------------
File W.mod has 15 rows and 20 columns
It contains the linear models resulting
from the separate multiple linear regression of each dependent variable
upon the set of explanatory variables
-------------------------------------------------------
File W.res has 15 rows and 20 columns
It contains (data - model) matrix
File :W.res
```

Les résidus sont sans tendance ni alternance. On y trouve par auto modélisation d'un tableau homogène qu'une grande partie de la variance résiduelle est encore liée à un modèle commun :

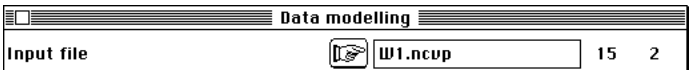


Pour obtenir cette figure renommer le fichier W.res en W_res et utiliser PCA : Non centred PCA :

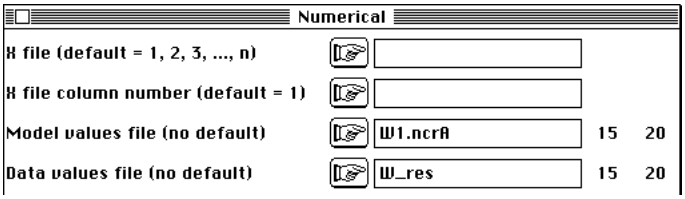




Reconstituer le tableau avec DDUtil : Data modelling :



Utiliser alors CurveModels pour représenter le résultat :



La régression sur des explicatives non corrélées est donc simple et très efficace. Les variables mesurées sont automatiquement peu ou prou corrélées entre elles, mais les variables de synthèses que sont les coordonnées factorielles issues d'une analyse de base ne le sont pas. L'intérêt de ce type de variables réside dans le fait qu'ajouter une explicative dans une prévision ne modifie pas le modèle précédent. On l'a utilisé sur des vecteurs propres de voisinage dans deux conditions expérimentales opposées.

La régression linéaire multiple utilisant des coordonnées factorielles est appelées régression sur composantes[10] et a d'abord été introduite sur des tableaux d'explicatives traités en ACP. Elle a été étendue en écologie au cas de l'AFC [11]. On l'utilisera avec le module OrthoVar qui vient d'être décrit.

Références

[1] Ter Braak, C.J.F., Juggins, S., Birks, H.J.B. & Voet, H. Van der. (1993) Weighted averaging partial least squares regression (WA-PLS): definition and comparison with other methods for species-environment calibration. In : *Multivariate Environmental Statistics*. Patil, G.P. & Rao, C.R. (Eds.) North-Holland, Amsterdam. in press.

[2] Shelford V.E. (1911) Ecological succession: stream fishes and the method of physiographic analysis. *Biol. Bull. (Woods Hole)* 21, 9-34.

[3] Whittaker, R.H., Levin, S.A. & Root, R.B. (1973) Niche, habitat and ecotope. *American Naturalist* : 107, 321-338.

[4] Cleveland, W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* : 74, 829-836.

Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983) *Graphical methods for data analysis*. Duxbury Press, Boston. 1-395.

Cleveland, W.S. & Devlin, S.J. (1988) Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* : 83, 596-610.

[5] Trouvilliez, J. (1988) *Contribution à l'étude des relations interspécifiques chez les oiseaux aquatiques. L'association entre le Grèbe à cou noir, Podiceps nigricollis, et la Mouette rieuse, Larus ridibundus, en période de nidification.* Thèse de doctorat, Université Lyon I. 1-228+annexes.

[6] Tisne-Agostini, D. (1988) Description par analyse en composantes principales de l'évolution de la production du clémentinier en association avec 12 types de porte-greffe. Rapport technique, DEA Analyse et modélisation des systèmes biologiques, Université Lyon 1. 1-26 + Annexes.

[7] INRA, 20230 San Nicolao, France.

[8] Méot, A., Chessel, D. & Sabatier, R. (1993) Opérateurs de voisinage et analyse des données spatio-temporelles. In : *Biométrie et Environnement. Lebreton*, J.D. & Asselain, B. (Eds.) Masson, Paris. 45-72.

[9] Thioulouse J., Chessel D & Champely S. (1995) Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environmental and Ecological Statistics* (in revision).

[10] Næs, T. (1984) Leverage and influence measures for principal component regression. *Chemometrics and Intelligent Laboratory Systems* : 5, 155-168.

[11] Roux, M. (1973) Estimation des paléoclimats d'après l'écologie des foraminifères [Paléoclimats]. *Les Cahiers de l'Analyse des Données* : 4, 61-79.