Régression multiple : principes et exemples d'application

Dominique Laffly
UMR 5 603 CNRS
Université de Pau et des Pays de l'Adour
Octobre 2006

Destiné à de futurs thématiciens, notamment géographes, le présent exposé n'a pas pour vocation de présenter la théorie de l'analyse des données par régression au sens statistique du terme. Pour cela nous renvoyons aux nombreux ouvrages rédigés par les statisticiens eux-mêmes. Le but recherché ici est de proposer des exemples concrets de traitement ayant fait appel à l'analyse par régression linéaire multiple selon différentes logiques *a priori* éloignées les unes des autres. Nous verrons successivement comment la méthode des régressions linéaires multiples permet :

- d'analyser les liens entre une variable dépendante quantitative à expliquer et plusieurs variables quantitatives explicatives indépendantes comme on l'admet généralement ;
- de déterminer les équations d'un ajustement polynomial non-linéaire pour l'analyse des liens entre deux variables quantitatives ;
- de déterminer les équations de surfaces de tendances ;
- d'analyser la rugosité du relief ;
- de déterminer les équations polynomiales d'un modèle de correction géométrique applicable à des vecteurs et/ou des données raster.

1. RÉGRESSION LINÉAIRE : LES PRINCIPES

L'analyse descriptive des données repose sur une démarche en plusieurs étapes. On définit tout d'abord les caractéristiques des variables prises une à une (analyse univariée ou tri à plat), puis on observe les liens qui les caractérisent deux par deux (analyse bivariée ou tri

croisée) pour finir par l'observation des structures multiples liant plusieurs variables (analyse multivariée). On distingue alors deux familles principales, la première consiste à observer les liens unissant une variable avec plusieurs autres (1 → *n*), la seconde considère simultanément les structures multiples liant différentes variables (*n* → *n*, analyse factorielle). Selon la nature des variables retenues les méthodes de calcul seront différentes mais la logique reste la même. L'analyse par régression linéaire multiple est une des solutions qui existe pour observer les liens entre une variable quantitative dépendante et *n* variables quantitatives indépendantes.

Toutes méthodes faisant appel aux régressions reposent sur l'acceptation des hypothèses fondatrices de la statistique paramétrique[1] et la notion d'ajustement par les moindres carrés. La moyenne arithmétique d'une variable est par conséquent considérée comme un centre de gravité et la notion des moindres carrés consiste à minimiser la somme des résidus élevés à la puissance deux entre la valeur observée et celle extrapolée.

1.1. Régression linéaire simple

Un exemple simple d'ajustement par les moindres carrés est donné par l'analyse bivariée de variables quantitatives qui peut se simplifier par le calcul des variances et de la covariance des deux variables X et Y retenues.

La variance répond à la formule suivante :

$$VarX = \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

où : *n*, nombre d'individus

x_i, valeur de la variable *x* pour l'individu *i*

$\bar{x}$, moyenne arithmétique de la variable *x*

[1] Pour simplifier à l'extrême, la statistique paramétrique repose sur l'hypothèse que les données sont des variables indépendantes distribuées selon une loi normale.

La covariance considère les variations communes des deux variables selon la formule :

$$CovXY = \frac{1}{N}\sum_{i=1}^{n}\left(x_i - \bar{x}\right) * \left(y_i - \bar{y}\right)$$

où : n, nombre d'individus

x_i, valeur de la variable x pour l'individu i

$\bar{x}$, moyenne arithmétique de la variable x

y_i, valeur de la variable x pour l'individu i

$\bar{y}$, moyenne arithmétique de la variable y

Enfin, le coefficient de corrélation est donné par la formule :

$$Ceof.cor = \frac{CovXY}{\sqrt{VarX} * \sqrt{VarY}}$$

Le coefficient de corrélation correspond au cosinus de l'angle formé entre deux droites de régression se croisant aux coordonnées des moyennes arithmétiques des deux variables observées (centre de gravité supposé). On définit donc deux droites répondant chacune à une équation affine :

$$X' = a1Y + b1$$

et

$$Y' = a2X + b2$$

X' et *Y'* étant les valeurs estimées à partir des valeurs observées *X* et *Y*.

Dans le cas de l'analyse bivariée, les coefficients des équations sont facilement donnés par :

$$a1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

$$a2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

$$b1 = \bar{y} - a1\bar{x}$$

$$b2 = \bar{x} - a2\bar{y}$$

Prenons comme exemple la matrice théorique suivante (table A1) :

id	X	Y	X'	Y'	X-moyX	Y-moyY	(X-moyX)²	(Y-moyY)²	(X-moyX)(Y-moyY)
1	2	18	1.847222222	13.95157895	-4.777777778	8.333333333	22.82716049	69.44444444	-39.81481481
2	3	15	3.622222222	13.05473684	-3.777777778	5.333333333	14.27160494	28.44444444	-20.14814815
3	4	12	5.397222222	12.15789474	-2.777777778	2.333333333	7.716049383	5.444444444	-6.481481481
4	5	9	7.172222222	11.26105263	-1.777777778	-0.666666667	3.160493827	0.444444444	1.185185185
5	6	6	8.947222222	10.36421053	-0.777777778	-3.666666667	0.604938272	13.44444444	2.851851852
6	8	5	9.538888889	8.570526316	1.222222222	-4.666666667	1.49382716	21.77777778	-5.703703704
7	10	6	8.947222222	6.776842105	3.222222222	-3.666666667	10.38271605	13.44444444	-11.81481481
8	11	7	8.355555556	5.88	4.222222222	-2.666666667	17.82716049	7.111111111	-11.25925926
9	12	9	7.172222222	4.983157895	5.222222222	-0.666666667	27.27160494	0.444444444	-3.481481481

Table A1 : Exemple théorique

Le coefficient de corrélation est de -*0.72844463*, les équations sont :

$Y' = -0.8968X + 15.745$ (en jaune)

et

$X' = -0.5917Y + 12.497$ (en magenta)

La somme des carrés des écarts entre les valeurs observées et celles théoriques est ici minimale pour les deux droites de régression, ce qui correspond à l'ajustement par les moindres carrés. Notons que ces écarts sont appelés résidus et qu'ils sont perpendiculaires (c'est-à-dire indépendants d'un point de vue mathématique) à l'axe de la variable explicative dont les valeurs ne changent pas par définition (figure A8).

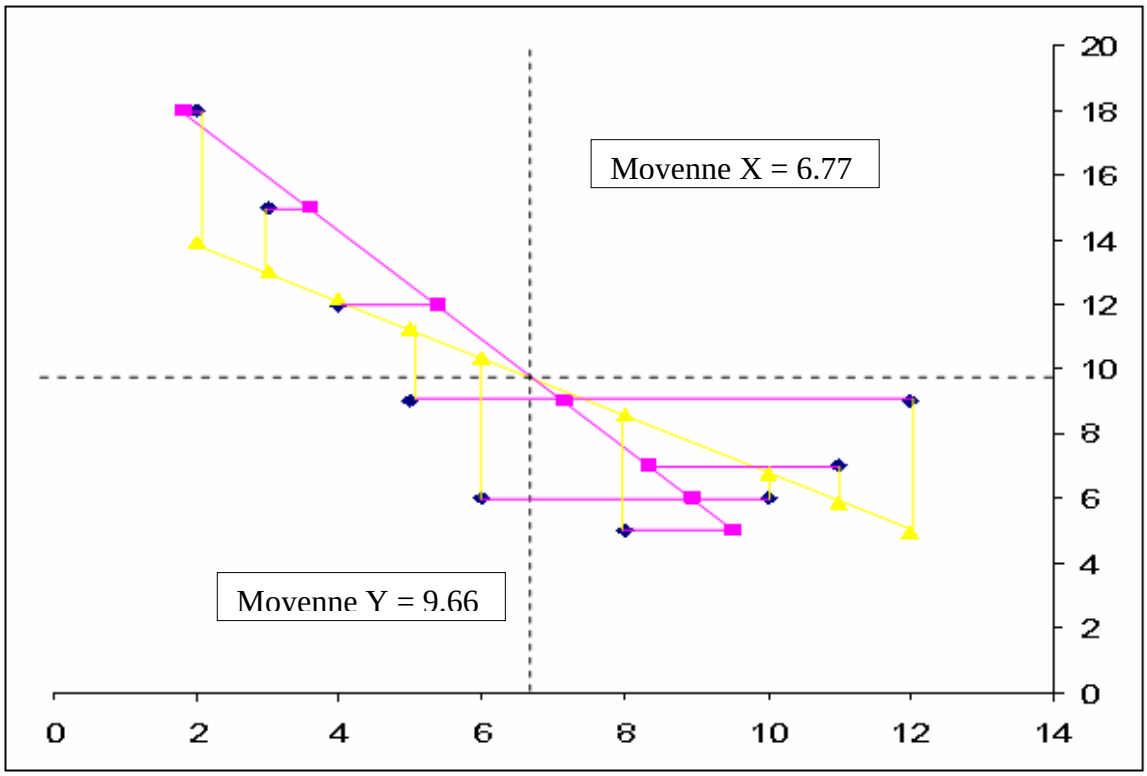

Figure A8 : Les deux droites de régression et le coefficient de corrélation

1.2. Régression linéaire multiple

L'exemple développé à partir de deux variables permet de comprendre la logique de la théorie de la régression mais il ne peut être généralisé de la sorte aux régressions multiples. Le système à deux équations à deux inconnus présenté se résolvait facilement comme on l'a vu. Les équations se compliquent avec plusieurs régresseurs, deux méthodes distinctes permettent de résoudre les équations. La première repose sur la connaissance des coefficients de corrélation linéaire simple de toutes les paires de variables entre elles, de la moyenne arithmétique et des écarts-types de toutes les variables. La seconde repose sur des calculs matriciels.

1.2.1. Les étapes de calcul fondé les variables descriptives

Soit un ensemble de *p* variable où la *p-ième* variable est la variable indépendante. Toutes les variables sont au préalable centrées-réduites. Soit r_{12}, r_{13}… $\mathbf{r}_{pp}$ les coefficients de corrélations linéaires des paires de variables et s_1, s_2, …, s_p les écarts-types.

Prenons un exemple avec $p = 4$ soit 3 variables dépendantes. Dans un premier temps on calcule les coefficients de régression linéaire a'_1, a'_2, a'_3 en résolvant un système de *p-1* équations à *p-1* inconnues :

$$r_{1p} = a'_1 + r_{12}a'_2 + r_{13}a'_3$$

$$r_{2p} = a'_2 + r_{21}a'_1 + r_{23}a'_3$$

$$r_{3p} = a'_3 + r_{31}a'_1 + r_{32}a'_3$$

Pour résoudre ce système on procède par substitutions successives :

$$a'_1 = r_{1p} - r_{12}a'_2 + r_{13}a'_3$$

d'où

$$r_{2p} = a'_2 + (r_{21} * (r_{1p} - r_{12}a'_2 + r_{13}a'_3)) + r_{23}a'_3$$

$$a'_2 = r_{2p} - r_{21}a'_1 + r_{23}a'_3$$

$$a'_3 = r_{3p} - r_{31}a'_2 + r_{32}a'_3$$

Connaissant désormais les coefficients de régression on détermine ceux des variables brutes :

$$a_1 = a_1' \frac{S_y}{S_{x1}} \; ; \; a2 = a_2' \frac{S_y}{S_{x2}} \text{ et } a3 = a_3' \frac{S_y}{S_{x3}}$$

Enfin, la constante d'ajustement est donnée en résolvant l'équation pour la coordonnée à l'origine :

$$\varepsilon = \bar{y} - a_1 \bar{x}_1 - a_2 \bar{x}_2 - a_3 \bar{x}_3$$

Le coefficient de détermination multiple est donné par :

$$R^2 = \sum_{j=1}^{p-1} a_j' r_{jp}$$

Prenons garde au fait que ce coefficient – dont les a'_{p-1} constituent en quelque sorte la contribution – croît avec le nombre de variable. Par conséquent, ce comportement déterministe lié aux propriétés des variables aléatoires doit être compenser, on calcule alors le coefficient ajusté :

$$R^2 ajusté = 1 - \frac{(n-1)}{n-(p-1)-1}(1-R^2)$$

Où : n : nombre d'individus

On peut également résoudre le système d'équations en prenant comme principe l'justement par les moindres carrés (Chadule) :

$$\sum_{i=1}^{n} \varepsilon_i^2 \text{ min}$$

Où : ε : variance résiduelle

Les coefficients a_j sont alors extraits des équations :

$$Cov_{p,1} = a_1 Var_1 + a_2 Cov_{1,2} + ... + a_{p-1} Cov_{1,p-1}$$
$$Cov_{p,2} = a_1 Cov_{2,1} + a_2 Var_2 + ... + a_{p-1} Cov_{2,p-1}$$
…
$$Cov_{p,p-1} = a_1 Covp-1,_1 + a_2 Cov_{1,2} + ... + a_{p-1} Var_{p-1}$$

Les *p-1* coefficients sont ensuite obtenus par résolution du système. Avec deux variables explicatives X_1 et X_2 et une variable à expliquer Y on a par exemple :

$$a_1 = \frac{\left(Var_{X_2} * Cov_{YX_1}\right) - \left(Cov_{YX_2} * Cov_{X_1X_2}\right)}{\left(Var_{X_1} * Var_{X_2}\right) - Cov^2_{X_1X_2}} = \frac{\sigma_Y * \left(r_{YX_1} - \left(r_{YX_2} * r_{X_1X_2}\right)\right)}{\sigma_{X_1} * \left(1 - r^2_{X_1X_2}\right)}$$

$$a_2 = \frac{\left(Var_{X_1} * Cov_{Y_{21}}\right) - \left(Cov_{YX_1} * Cov_{X_1X_2}\right)}{\left(Var_{X_1} * Var_{X_2}\right) - Cov^2_{X_1X_2}} = \frac{\sigma_Y * \left(r_{YX_2} - \left(r_{YX_1} * r_{X_1X_2}\right)\right)}{\sigma_{X_{21}} * \left(1 - r^2_{X_1X_2}\right)}$$

Le coefficient de corrélation multiple est alors donnée par :

$$R_{Y,X_1X_2} = \sqrt{\frac{\left(r^2_{YX_1} + r^2_{YX2} - 2\left(r_{YX_1} * r_{YX_2} * r_{X_1X_2}\right)\right)}{1 - r^2_{X_1X_2}}} = r_{YY'}$$

1.2.2. La notation matricielle

L'équation de type :

$$\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x}_2 + \beta_2 1\vec{x}_2 + \varepsilon$$

est donnée sous forma matricielle par :

$$y = X\beta + \varepsilon$$

Où :

$$y = \begin{matrix} y_1 \\ y_2 \\ \dots \\ y_{n-1} \\ y_n \end{matrix}, X = \begin{matrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ \dots & \dots & \dots \\ 1 & x_{1,n-1} & x_{2,n-1} \\ 1 & x_{1,n} & x_{2,n} \end{matrix}, \beta = \begin{matrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{matrix}, \varepsilon = \begin{matrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{matrix}$$

Il s'agit dès lors de calculer le vecteur des estimateurs $\hat{\beta}$ défini par l'égalité suivante :

$$\hat{\beta} = (X * X')^{-1} X' y$$

En notation matricielle X' signifie la matrice X transposée et X^{-1} la matrice inverse.

Dans l'exemple qui suit nous réalisons une régression multiple pour expliquer la hauteur de neige en fonction de l'altitude, de la rugosité, de la pente, de l'orientation, de la latitude et de la longitude (table A2).

H_NEIGE	vecteur	altitude	rugosite	pente	orient.	lat	long.
95	1	2768	252	22	324	8760219	438465.0625
150	1	4108	333	29	308	8760195	438474.0625
4	1	4045	62	5	249	8760168	438480.0625
0	1	4572	85	8	14	8760135	438489.0625
0	1	4614	115	10	63	8760105	438495.0625
80	1	4321	176	16	130	8760072	438498.0625
95	1	3886	72	6	199	8760039	438504.0625
20	1	4206	57	5	32	8760012	438507.0625
90	1	4192	266	23	197	8759985	438513.0625
10	1	4051	69	6	113	8759955	438519.0625
10	1	3746	62	5	149	8759922	438519.0625
50	1	3789	42	3	218	8759895	438525.0625
45	1	3771	44	4	53	8759865	438531.0625
60	1	3796	48	4	101	8759838	438534.0625
55	1	3885	77	7	332	8759811	438537.0625
3	1	4295	113	10	18	8759787	438540.0625
33	1	4467	147	13	50	8759760	438546.0625

0	1	4764	12	1	276	8759730	438552.0625
35	1	4313	38	3	350	8759703	438552.0625
45	1	4387	40	3	46	8759673	438558.0625

Table A2 : Hauteur de neige et variables environnementales

Le produit X'X donne :

20.0000	81976.0000	2110.0000	183.0000	3222.0000	175198869.0000	8770339.2500
81976.0000	339594498.0000	8487334.0000	736618.0000	12861325.0000	718104679425.0000	35947950323.5000
2110.0000	8487334.0000	366956.0000	32036.0000	386290.0000	18483638688.0000	925244282.8750
183.0000	736618.0000	32036.0000	2799.0000	33323.0000	1603083666.0000	80246258.4375
3222.0000	12861325.0000	386290.0000	33323.0000	771684.0000	28224580695.0000	1412891754.3750
175198869.0000	718104679425.0000	18483638688.0000	1603083666.0000	28224580695.0000	1534732185500860.0000	76827675778567.3000
8770339.2500	35947950323.5000	925244282.8750	80246258.4375	1412891754.3750	76827675778567.3000	3845942542298.3300

D'où $(X'X)^{-1}$:

42548515331.8374	73.5283	-569.7835	4096.6641	-164.4807	-3668.8247	-23739.2652
73.5284	0.0000	0.0000	-0.0001	0.0000	0.0000	0.0000
-569.7830	0.0000	0.0047	-0.0535	0.0000	0.0001	0.0003
4096.6572	-0.0001	-0.0535	0.6061	0.0005	-0.0004	-0.0014
-164.4807	0.0000	0.0000	0.0005	0.0000	0.0000	0.0001
-3668.8247	0.0000	0.0001	-0.0004	0.0000	0.0003	0.0020
-23739.2657	0.0000	0.0003	-0.0014	0.0001	0.0020	0.0133

Le produit X'X est donnée par la formule :

$$a_{i,j} = \sum_{k=1}^{n} b_{i,k} c_{k,j}$$

Où : a : matrice résultat ;

b et c : matrices initiales ;

i : lignes ;

j : colonnes.

Le produit d'une matrice de *k* lignes et *l* colonnes par une matrices de *l* lignes par *k* colonnes donne une matrice carrée de *k* lignes et colonnes. D'où la matrice carrée suivante :

20.0000	81976.0000	2110.0000	183.0000	3222.0000	175198869.0000	8770339.2500
81976.0000	339594498.0000	8487334.0000	736618.0000	12861325.0000	718104679425.0000	35947950323.5000
2110.0000	8487334.0000	366956.0000	32036.0000	386290.0000	18483638688.0000	925244282.8750
183.0000	736618.0000	32036.0000	2799.0000	33323.0000	1603083666.0000	80246258.4375
3222.0000	12861325.0000	386290.0000	33323.0000	771684.0000	28224580695.0000	1412891754.3750
175198869.0000	718104679425.0000	18483638688.0000	1603083666.0000	28224580695.0000	1534732185500860.0000	76827675778567.3000
8770339.2500	35947950323.5000	925244282.8750	80246258.4375	1412891754.3750	76827675778567.3000	3845942542298.3300

L'inversion d'une matrice fait appel à des notions de calculs matriciels poussés que nous ne développerons pas ici. Retenons qu'en théorie toute matrice dont le déterminant est non nul peut être inversée (règle de Cramer). D'où dans notre exemple $(X'X)^{-1}$:

42548515331.8374	73.5283	-569.7835	4096.6641	-164.4807	-3668.8247	-23739.2652
73.5284	0.0000	0.0000	-0.0001	0.0000	0.0000	0.0000
-569.7830	0.0000	0.0047	-0.0535	0.0000	0.0001	0.0003
4096.6572	-0.0001	-0.0535	0.6061	0.0005	-0.0004	-0.0014
-164.4807	0.0000	0.0000	0.0005	0.0000	0.0000	0.0001
-3668.8247	0.0000	0.0001	-0.0004	0.0000	0.0003	0.0020
-23739.2657	0.0000	0.0003	-0.0014	0.0001	0.0020	0.0133

Et X'y :

880
3458806
140963
12244
181900
7708792743
385887448

Donc $(X'X^{)-1}X'y$ donne les termes de l'équation multiple :

Constante : -6111180.498

Altitude : -0.03526

Rugosité : 1.0379

Pente : -7.6228

Orientation : 0.0907

Latitude : 0.5191

Longitude : 3.6401

2. EXEMPLES D'APPLICATION

L'utilisation des régressions multiples dépasse largement le cadre classique de l'explication d'une variable dépendante à partir de *n* variables indépendantes comme on l'admet généralement. Nous verrons tout d'abord un exemple appliqué à l'analyse du trachome[2] en fonction de paramètres biogéographiques pour illustrer cette approche classique. Trois autres exemples nous permettront d'aller plus en avant dans l'application des régressions multiples : l'ajustement non linéaire en analyse bivariée ; l'analyse par surfaces de tendance d'un phénomène géographique et la définition des équations d'un modèle de correction géométrique.

2.1. Indicateurs environnementaux et Trachome

Le trachome est une maladie contagieuse qui se transmet d'enfant à enfant ou de mère à enfant. L'infection se manifeste dès la première année et la prévalence augmente très rapidement pour atteindre un maximum qui serait d'autant plus précoce que le niveau de l'endémie est élevé. La prévalence du trachome actif diminue ensuite progressivement et laisse place à des lésions cicatricielles dont la fréquence augmente avec l'âge. Il n'y a pas de différence de prévalence selon le sexe significative dans l'enfance, par contre à l'âge adulte les femmes sont plus fréquemment atteintes du fait des contacts avec les enfants, elles présenteront par la suite plus fréquemment un entropion trichiasis que les hommes.

Le trachome actif est caractérisé par une inflammation de la conjonctive tarsale supérieure avec envahissement de la cornée par un voile vasculaire (pannus). Ce stade inflammatoire représente la phase contagieuse de la maladie. L'inflammation trachomateuse en milieu hyper-endémique persistera quelques années avant d'évoluer vers la cicatrisation qui pourra se faire selon deux modalités :

- soit l’infection est restée modérée et l’évolution se fera vers la guérison spontanée au prix de quelques cicatrices conjonctivales minimes sans conséquence fonctionnelles : c’est le trachome cicatriciel bénin.

- soit l’inflammation conjonctivale a été intense et prolongée : la cicatrisation pourra alors dépasser son but et entraîner une fibrose rétractile de la paupière supérieure. Il s’agit alors d’un trachome cicatriciel grave susceptible d’aboutir à une déformation du tarse avec déviation des cils vers la cornée réalisant un entropion trichiasis. Le frottement des cils à chaque clignement entretien une érosion cornéenne particulièrement douloureuse, favorisant une surinfection qui évoluera vers une cécité complète et irréversible par opacification de la cornée. Une fois les lésions cicatricielles constituées, le seul moyen d'améliorer le pronostic et si possible d’empêcher la cécité est la chirurgie du trichiasis : les techniques chirurgicales sont relativement efficaces et sûres, mais elles sont insuffisamment diffusées et utilisées.

C’est la durée et surtout l’intensité de l’inflammation trachomateuse qui déterminent le risque de l’évolution vers la cécité. Cette intensité est conditionnée par deux facteurs : les surinfections bactériennes et les réinfections. La plus grande gravité des réinfections est expliquée par un mécanisme combiné d’hypersensibilité et d’auto-immunité.

Un certain nombre de facteurs de risque associés au trachome ont été identifiés. Ces facteurs sont individuels, comportementaux, sociaux et aussi environnementaux. C'est ainsi que la difficulté d'accès à l'eau, l'accumulation d'ordures, la proximité avec le bétail et la pullulation des mouches favorise la survenue d'un trachome.

L’influence de la géographie et du climat est évoquée depuis longtemps dans le complexe pathogène du trachome. En zone intertropicale sèche, la diminution de l’humidité atmosphérique dessèche les muqueuses conjonctivales et favoriserait l’infection par les chlamydia. Les poussières pourraient jouer un rôle non négligeable en agressant la conjonctive et la cornée. Par ailleurs en hiver, le froid nocturne augmente la promiscuité dans les chambres et favoriserait la circulation interhumaine du germe. Dans une étude épidémiologique Salim rapporte qu’au Soudan la prévalence du trachome actif est inversement corrélée avec la pluviométrie et l’hygrométrie. Nous avons aussi observé dans l'enquête nationale réalisée au Mali, que le trachome actif était plus fréquent dans les régions sèches du nord comme Gao ou Tombouctou.

L’exemple développé ci-dessous s’inscrit pleinement dans la quatrième partie de l’ouvrage. Il s’agit de déterminer des facteurs environnementaux du risque au Mali.. Les données sont issues d’une enquête biomédicale réalisée auprès de 11 000 personnes en Afrique de l’ouest. Elles sont confrontés à différentes variables environnementales susceptibles d’être liées à la maladie : latitude (LAT), longitude (LONG), pluviométrie (PLUVIO), température moyenne annuelle (TMOY) et hygrométrie (HYGRO) à partir du fichier des villages. Ces variables ont été récupérées dans le fichier des individus femmes. Les données manquantes ont été extrapolées par régression linéaire multiple avec les variables LAT et LONG. Les données sur le trachome sont quant à elles : trachome actif (TT), trachome ches les femmes (TF), trachome chez les enfants (TI) et trachome suspecté (TS). La figure A9 présente les cartes des variables environnementales retenues.

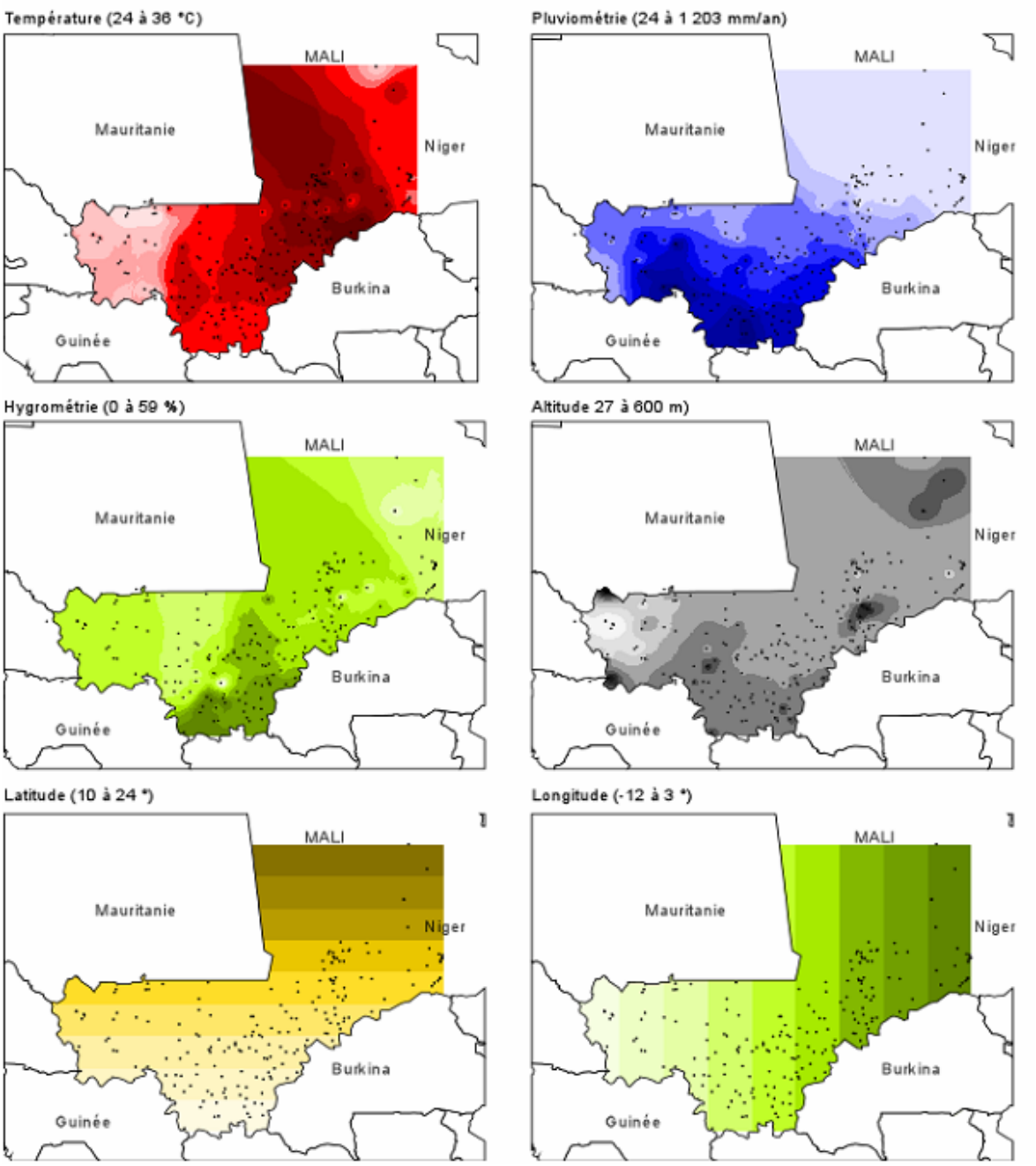

Figure A9 : Les variables environnementales

Par régression linéaire multiple, on calcule les degrés de liaisons entre les taux de prévalence entre la latitude, la longitude, la pluviométrie, l'altitude, la température moyenne et l'hygrométrie (table xx).

Coef. Cor.	LAT	LONG	Pluvio	Altitude	T moy	Hygro	PREVALTF	PREVALTI	PREVALTS	PREVALTT
LAT	1									
LONG	0.536995393	1								
Pluvio	-0.87120945	-0.51303925	1							
Altitude	-0.230234	0.084529641	0.26712518	1						
T moy	0.114032981	0.565620188	-0.20241602	0.189518905	1					
Hygro	-0.51904632	-0.09440315	0.443513521	0.148531308	0.238402184	1				
PREVALTF	0.222571057	0.023063302	-0.1408921	0.059441032	-0.15348681	-0.21528461	1			
PREVALTI	0.266251036	0.17773431	-0.20632332	0.006969171	-0.04865699	-0.09405617	0.621083998	1		
PREVALTS	-0.25314565	-0.32136359	0.30129919	0.216814719	0.053365569	0.222201689	0.166588961	0.298103035	1	
PREVALTT	-0.31651985	-0.33733686	0.32418421	0.24158158	-0.11456298	0.050142418	0.162682868	0.008269396	0.460840448	1

Coef. Det	LAT	LONG	Pluvio	Altitude	T moy	Hygro	PREVALTF	PREVALTI	PREVALTS	PREVALTT
LAT	1									
LONG	0.288364052	1								
Pluvio	0.160548815	0.328373985	1							
Altitude	0.053007693	0.007213044	0.071356183	1						
T moy	0.013003521	0.319926815	0.040973053	0.035940161	1					
Hygro	0.26940908	0.008912067	0.196751468	0.022061519	0.056835601	1				
PREVALTF	0.049537875	0.000531916	0.019850753	0.003533236	0.023558202	0.046341462	1			
PREVALTI	0.070889614	0.031589485	0.042569313	4.85693E-05	0.002367503	0.008846564	0.393234341	1		
PREVALTS	0.064082118	0.101166918	0.090781202	0.047006623	0.002850019	0.04931359	0.219515614	0.08886542	1	
PREVALTT	0.100184812	0.113796154	0.105095441	0.061298639	0.020898454	0.002514262	0.026465715	6.83829E-05	0.212313919	1

Table A3 : Tables des corrélations multiples

Les coefficients de corrélation multiple sont donnés par la table XX. A titre indicatif le cas 1 présente les valeurs pour une régression n'intégrant que la latitude et la longitude, le cas 2 intègre toutes les variables environnementales retenues. On constate que les corrélations obtenues sont toujours significativement plus élevées dans la cas 2.

	Cas 1	Cas 2
TT	0.37	0.45
TS	0.34	0.50
TI	0.27	0.32
TF	O.25	0.33

Table A4 : Taux de corrélation multiple

On peut dès lors envisager de produire des cartes de prédiction des taux de prévalence du trachome et des résidus connaissant les variables environnementales. Les coefficients des équations sont donnés par la matrice A5.

	TF (0.33)	*TI (0.32)*	*TS (0.50)*	*TT (0.45)*
Constante	14.05981086	-0.166492463	-121.0795444	2.083385988
LAT	3.615939928	1.223183048	2.6892253	-0.248426023
LONG	-0.190359364	0.314126994	-3.111756389	-0.283659613
PLUVIO	0.005563127	0.00153108	0.009886783	7.52743E-05
ALT	0.033674547	0.005407455	0.039777358	0.008239454
TMOY	-1.25696877	-0.484097825	1.983162114	0.018222353
HYGRO	-0.115784009	0.061047239	0.225234599	-0.027915423

Table A5 : Coefficients de régression multiple

D’où, par exemple, pour TT :

$$TT_{estimé} = (-0.248426023 * LAT) - (-0.283659613 * LONG) + (7.52743E-05 * PLUVIO) + (0.008239454 * ALT) + (0.018222353 * TMOY) - (0.027915423 * HYGRO) + 2.083385988$$

La figure A10 présente les cartes des valeurs estimées de prévalence de TT, TI, TF et TS. On constate bien évidemment des écarts entre la simulation et les valeurs mesurées dans les villages, le modèle n’expliquant que 40 % de la distribution. Une carte de taux de prévalence de TT obtenue par interpolation spatiale – inversement proportionnelle à la distance - est présentée en vis-à-vis de celle issue de la modélisation par régression multiple. Les deux documents sont très différents et l’on pourra retenir que ce n’est pas la proximité à un lieu caractérisé par de fort taux qui explique la répartition spatiale de TT.

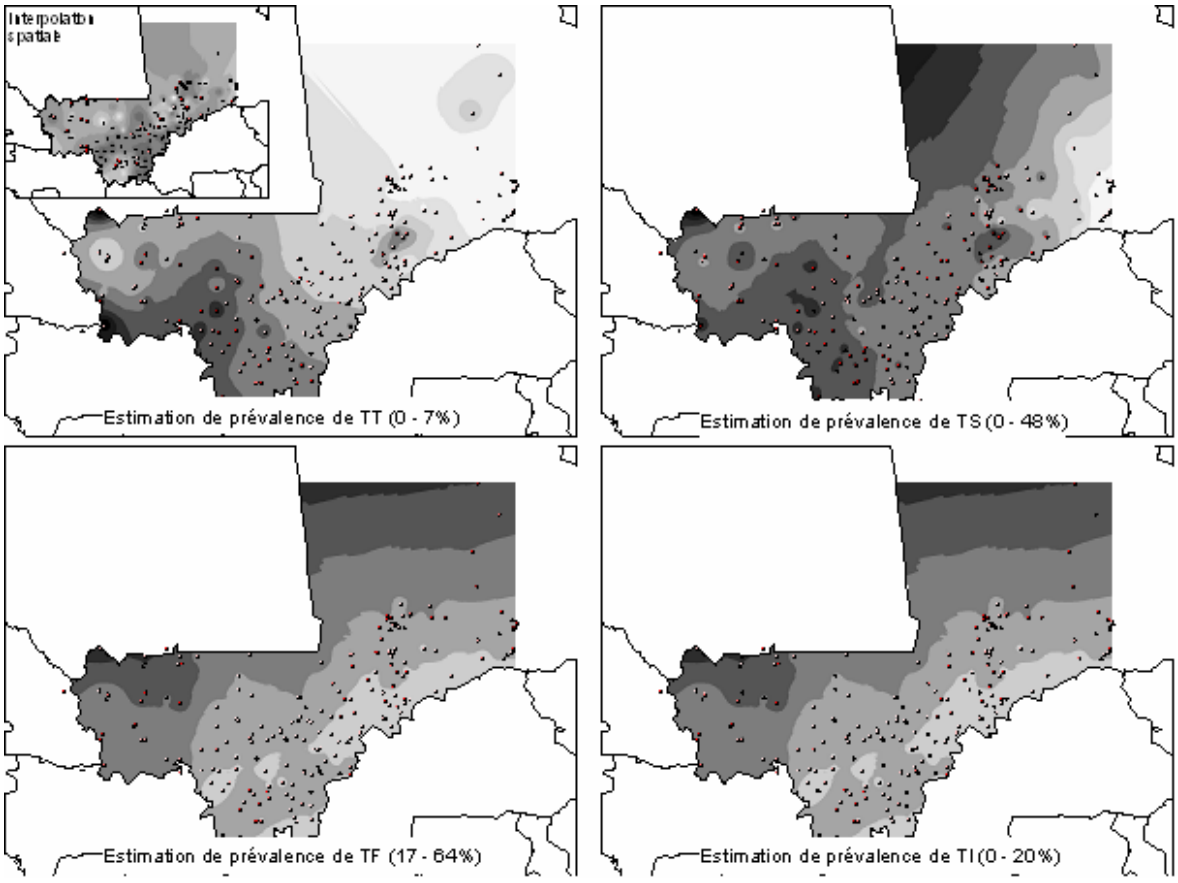

Figure A10 : Estimation des taux de prévalence par régression multiple

2.2. Ajustement non-linéaire et régression multiple

L'exemple théorique développé en introduction montrait un nuage de points distribués de manière non linéaire, d'où un faible coefficient de détermination (*0.53*) obtenu à partir d'un ajustement linéaire.

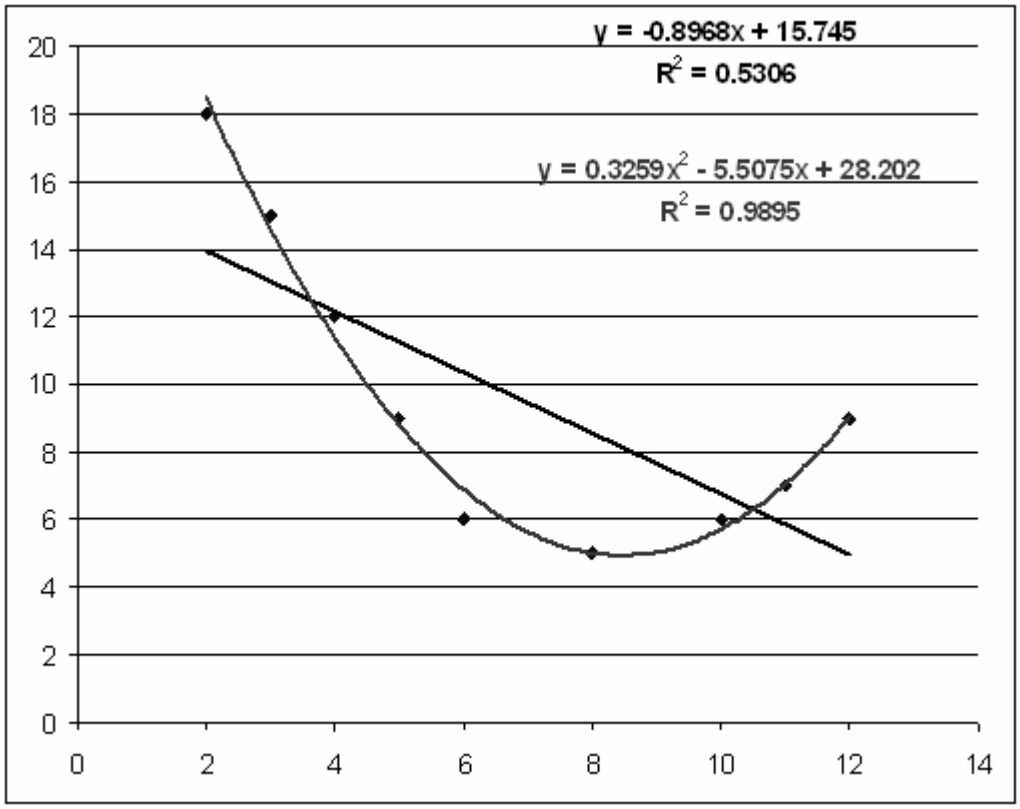

Figure A11 : Ajustement linéaire et non-linéaire d'ordre 2

On voit bien sur la figure A11 que les points répondent à une distribution qui s'aligne sur un morceau de parabole de type polynôme de degré 2 :

$$Y' = a_1 X^2 + a_2 X + b$$

Il s'agit en fait d'une régression linéaire multiple à partir d'une même variable X dont les termes sont élevés à la hauteur du degré du polynôme selon la formule générique :

$$Y' = a_1 X^1 + a_2 X^2 + ... + a_n X^n + \varepsilon$$

Dans l'exemple présenté plus haut, un ajustement d'un degré 2 permet d'obtenir un coefficient de détermination de l'ordre de *0.9895*. Lorsque les formes de la distribution sont plus complexes, on peut élever encore l'ordre du polynôme comme l'illustre la figure A12.

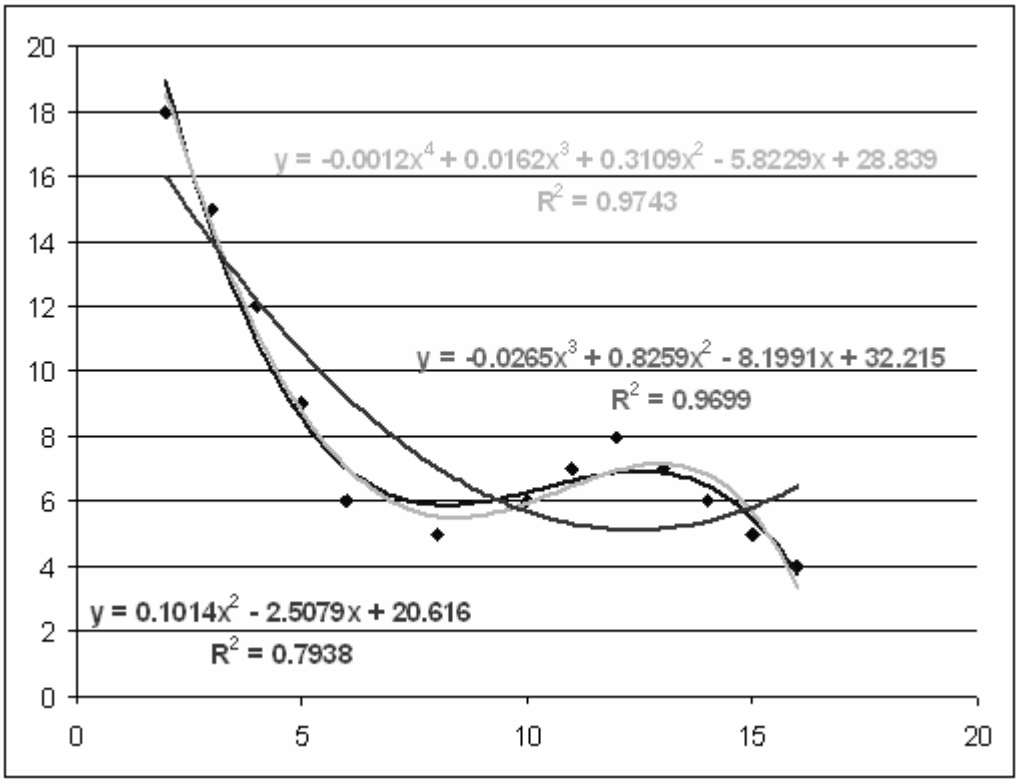

Figure A12 : Ajustements non linéaire d'ordre 3 à 5

Lorsque la distribution ne suit pas une loi polynomiale on peut observer les limites de l'ajustement comme l'illustre la figure ci-dessous. On peut alors avoir recours, selon la

forme, à une transformation logarithmique de la variable X pour donner une équation affine de type (figure A13) :

$$Y' = a.\ln(X) + b$$

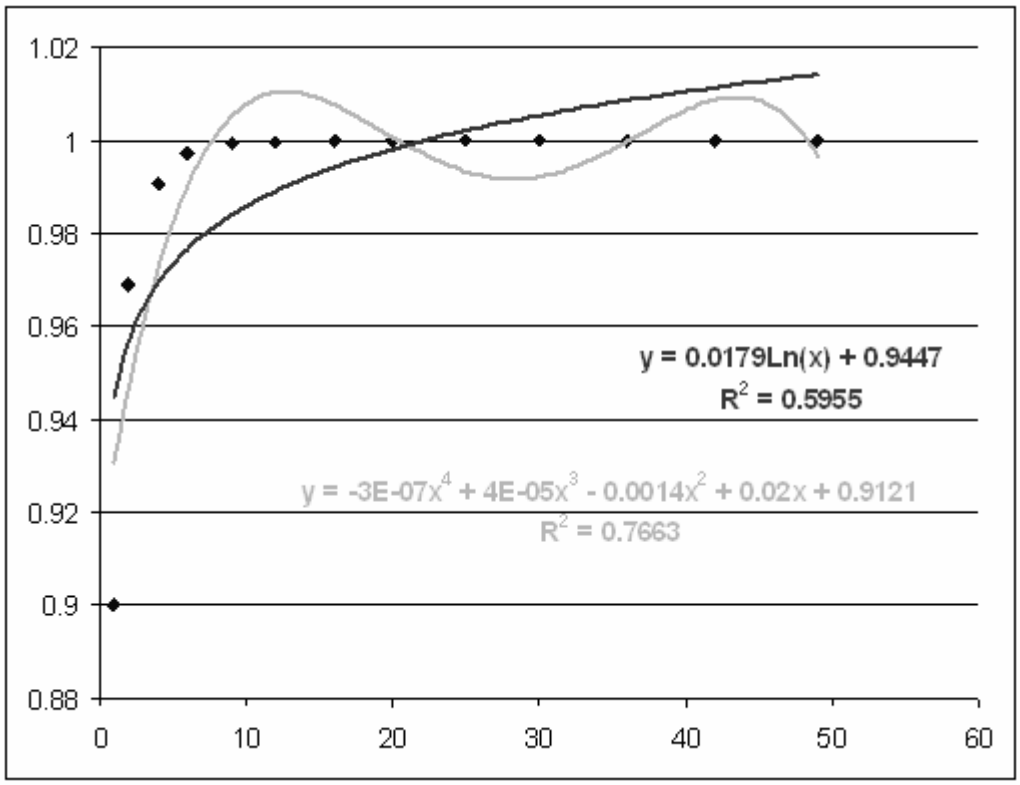

Figure A13 : Ajustements polynomial et logarithmique

2.3. Surfaces de tendances, régression multiple selon la latitude et la longitude

Lorsque les phénomènes étudiés sont fortement dépendants de leur position géographique on a recours aux surfaces de tendances pour extrapoler à l'ensemble de l'espace des valeurs initialement observées ponctuellement. Prenons un exemple théorique d'un phénomène marqué par un fort gradient sud ouest – nord est comme l'illustre les figures A14 et A15.

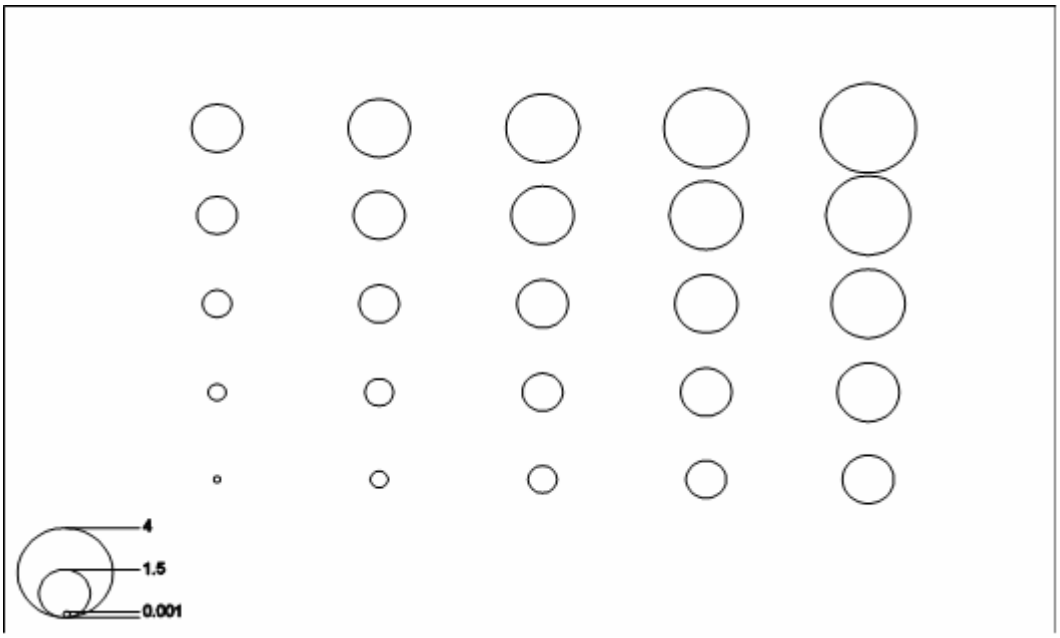

Figure A14 : Cartogramme d'un gradient

Une régression linéaire multiple avec comme variables indépendantes la latitude et la longitude nous donne ici un coefficient de détermination de 1 et une équation :

$$X' = 0.5Lat + 0.5Long - 0.5$$

Connaissant la latitude et la longitude on peut désormais extrapoler la variable X à tout l'espace géographique découpé en un maillage plus ou moins fin. On obtient alors un plan de régression ou surface de tendance d'ordre 1 comme l'illustre le schéma suivant :

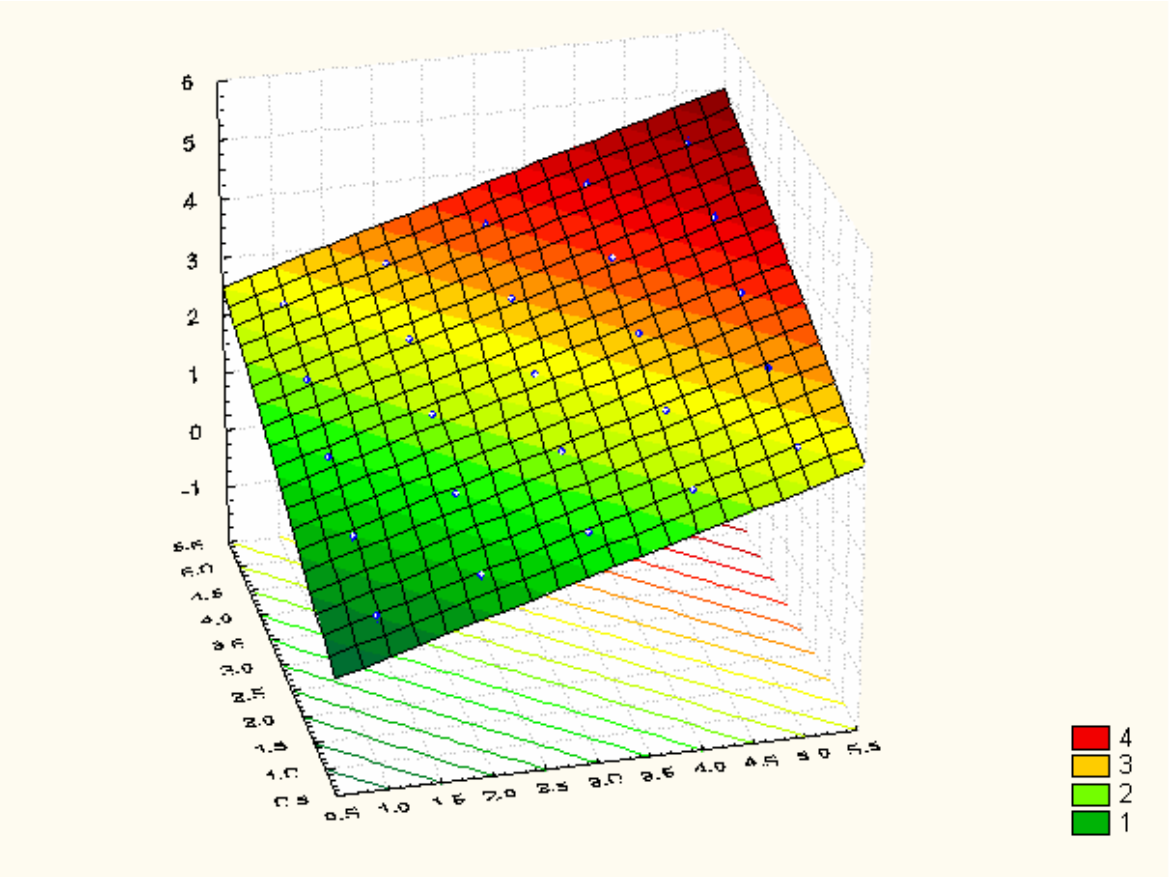

Figure A15 : Surface de tendance d'ordre 1 – plan de régression

On peut imaginer aisément une distribution géographique non linéaire d'un phénomène quelconque, une ondulation par exemple comme l'illustre les figures A16 et A17..

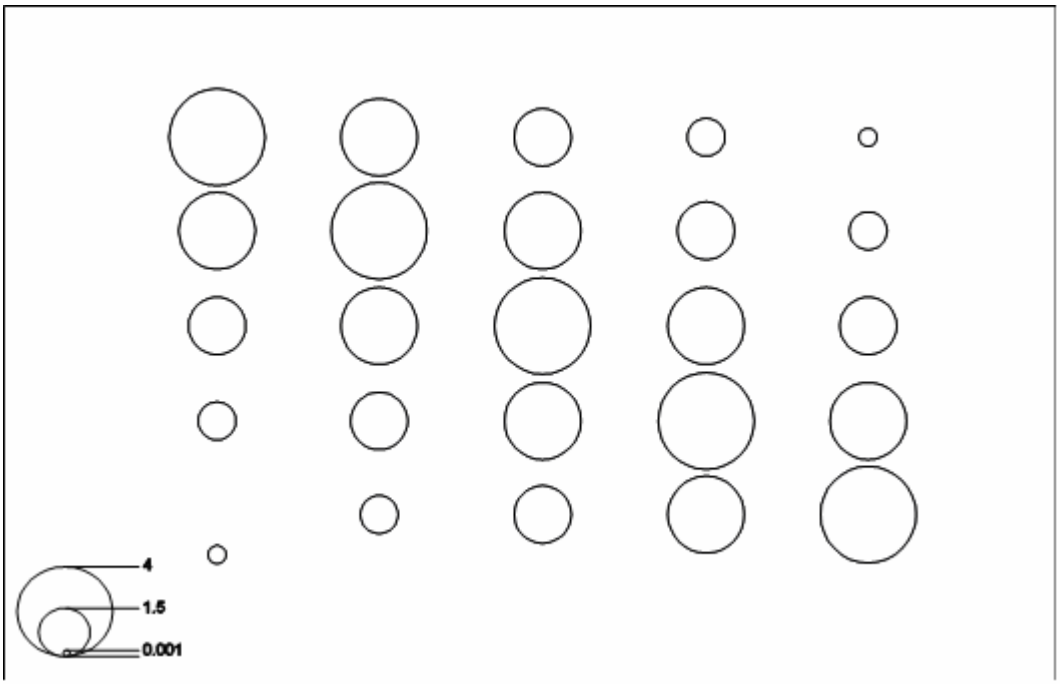

Figure A16 : Cartogramme d'une ondulation

L'ajustement d'un plan de régression ne serait alors pas du tout représentatif, dans notre exemple le coefficient de détermination est même nul. Comme on avait recours à des polynômes de degré *n* on peut envisager ici des surfaces polynomiales de degré *n*. Par exemple, l'équation d'une surface de degré 2 est :

$$X' = a_1 Lat + a_2 Long + a_3 Lat * Long + a_4 Lat^2 + a_5 Long^2 + \varepsilon$$

Soit présentement un coefficient de détermination de 0.95 et une équation :

$$X' = 1.38 Lat + 1.38 Long - 0.26 Lat * Long - 0.1 Lat^2 - 0.1 Long^2 - 2.04$$

On obtient alors la surface suivante :

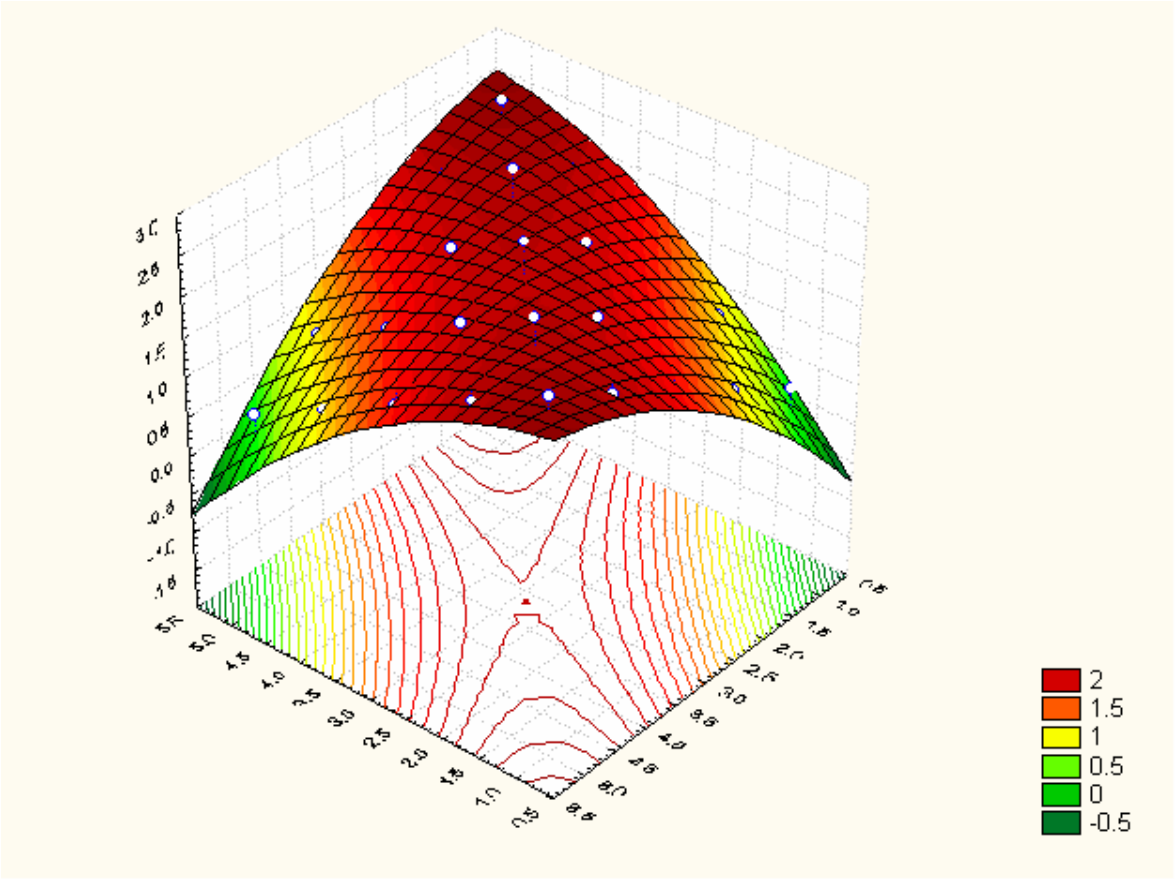

Figure A17 : Surface de tendance d'ordre 2

Pour des distributions plus complexes, on peut augmenter le degré du polynôme, par exemple une surface de degré 3 comme l'illustre les figures A18 et A19..

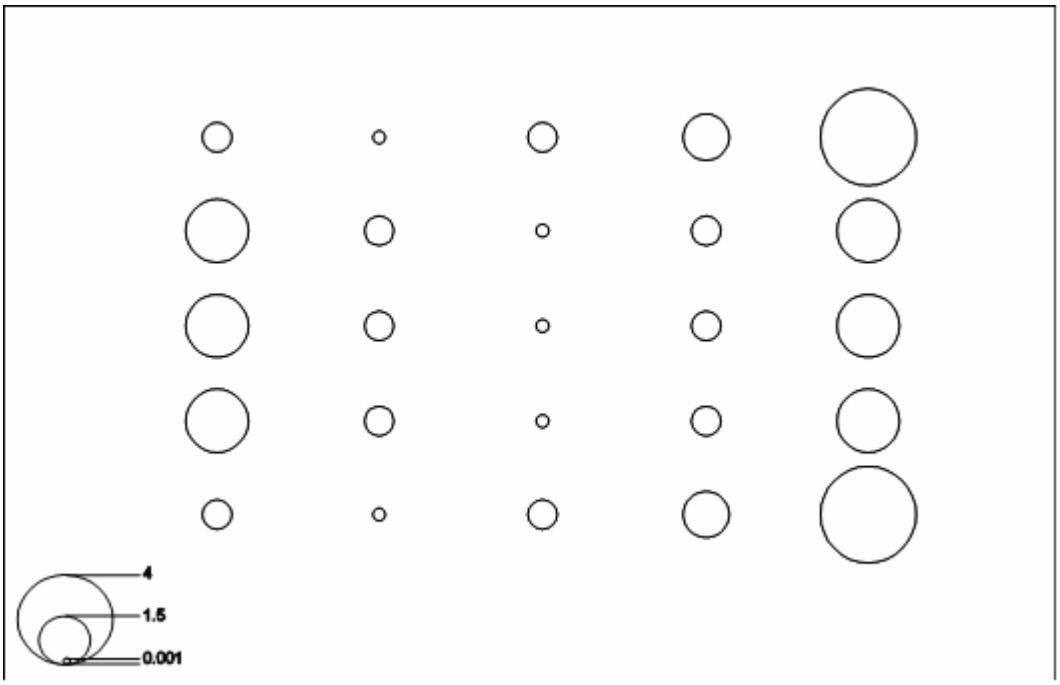

Figure A18 : Cartogramme d'une vague

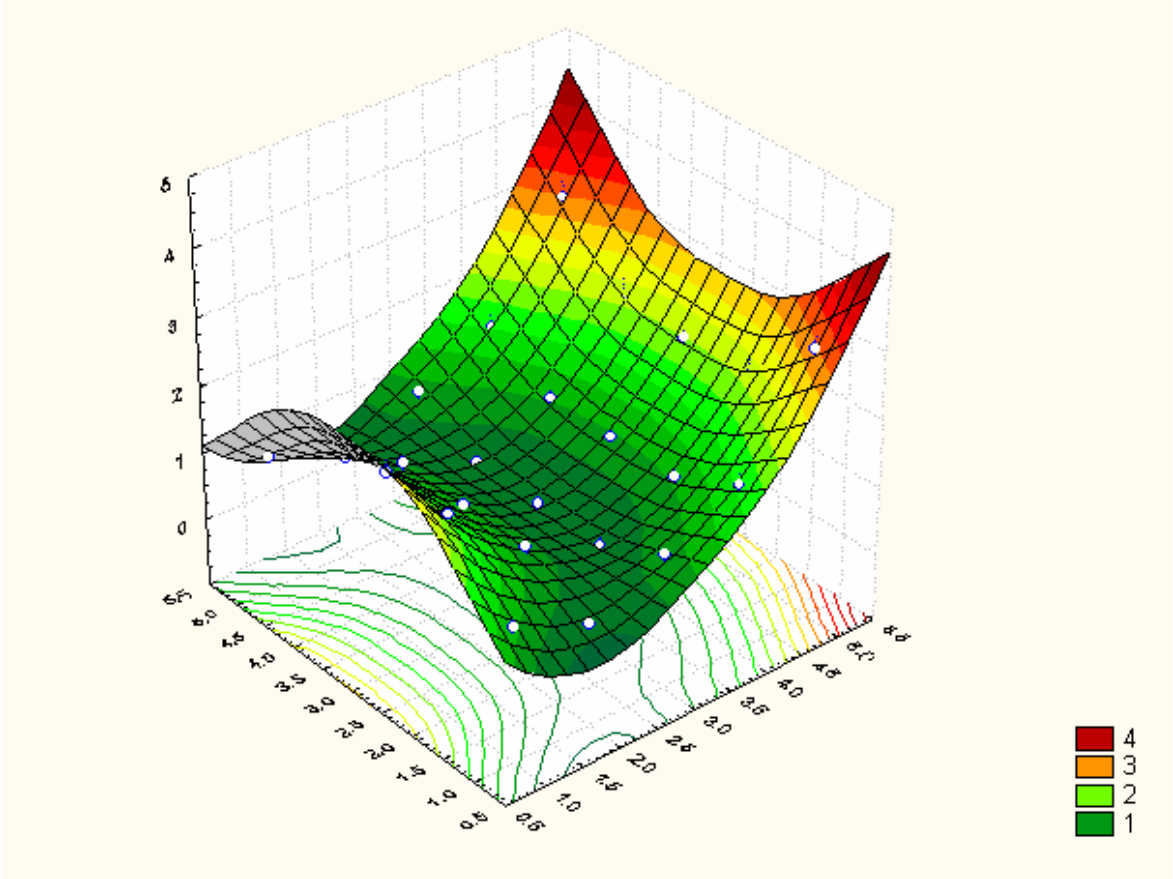

Figure A19 : Surface de tendance d'ordre 3

On atteint cependant rapidement des limites eut égard au nombre d'inconnus dans le système d'équations à résoudre. Celui-ci est donné par la formule :

$$Ninc = \sum_{i=1}^{t+1} i$$

t	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Ninc	*3*	*6*	*10*	*15*	*21*	*28*	*36*	*45*	*55*	*66*	*77*	*90*

où *t*, degré du polynôme.

L'équation polynomiale d'ordre *t* se généralise ensuite selon :

$$Z' = \sum_{i=0}^{t} \sum_{j=0}^{i} a_k x^{i-j} y^j$$

où

$$k = \frac{i.i + j}{2} + j$$

Pour un polynôme de degré 3, l'équation générique est de la forme :

$$Z' = a_1 X + a_2 Y + a_3 X^2 + a_4 XY + a_5 Y^2 + a_6 X^3 + a_7 X^2 Y + a_8 XY^2 + a_9 Y^3 + \varepsilon$$

Pour un polynôme de degré 4, elle devient :

$$Z' = a_1 X + a_2 Y + a_3 X^2 + a_4 XY + a_5 Y^2 + a_6 X^3 + a_7 X^2 Y + a_8 XY^2 + a_9 Y^3 + a_{10} X^4 + a_{11} X^3 Y^2 + a_{12} X^2 Y^2 + a_{13} X^2 Y^3 + a_{14} Y^4 + \varepsilon$$

Au-delà, le système d'équation à résoudre devient difficile à mettre en œuvre et, surtout, la surface extrapolée ne représente plus un phénomène géographique fortement dépendant de la latitude et de la longitude. Nous verrons dans un cours consacré au variogramme et covariogramme comment extrapoler des surfaces qui mettent en avant des cycles ou des phénomènes ponctuels.

2.4. Analyse du relief et indicateur de rugosité

Un cas particulier de d'application des surfaces de tendances, donc des régressions multiples, et celui de l'analyse de la « rugosité » du relief à partir d'un Modèle Numérique de Terrain (MNT). Un MNT est une base de données raster composée de *NC* colonnes et *NL* lignes dont les mailles renseignent sur l'altitude.

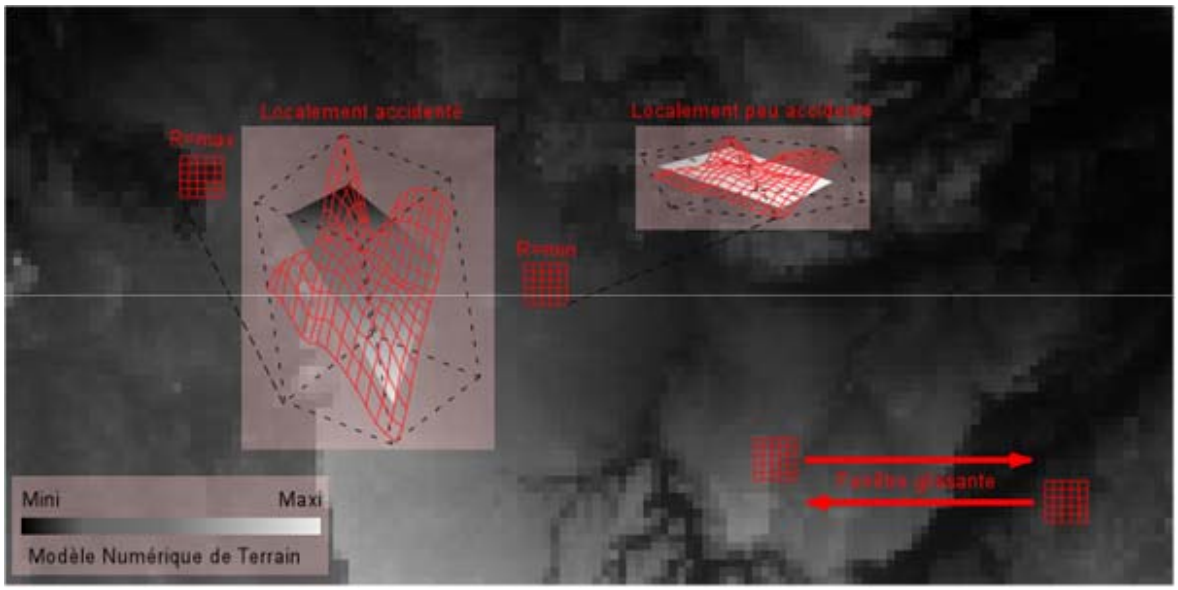

Figure A20 : Filtre de rugosité du relief

Afin d'obtenir une information synthétique sur les formes locales du relief on fait glisser sur chaque pixel de l'image une fenêtre de *n* mailles de côtés à l'intérieur de laquelle on ajsute une surface de tendance d'ordre 1 locale. La rugosité correspond à l'écart-type calculé sur les résidus entre les altitudes du MNT et celles de la surface de tendance. Plus la valeur de l'écart-type est élevée plus le relief est localement accidenté (creux, bosses,

crêtes, talwegs…) et inversement (plan d'un versant, d'une plaine…). La figure A20 illustre la méthode mise en œuvre.

Sans parler de la résolution du MNT, le choix de la taille de la fenêtre repose sur le niveau d'observation attendu par l'utilisateur. Une fenêtre de l'ordre de 3 à 9 pixels donnera des indicateurs locaux, ils seront plus globaux au-delà. La figure A21 présente une vue où différentes rugosités ont été calculées selon quatre tailles de fenêtres. Les teintes du rose au marron sont une gradation des valeurs de rugosité, respectivement de la plus faible à la plus forte.

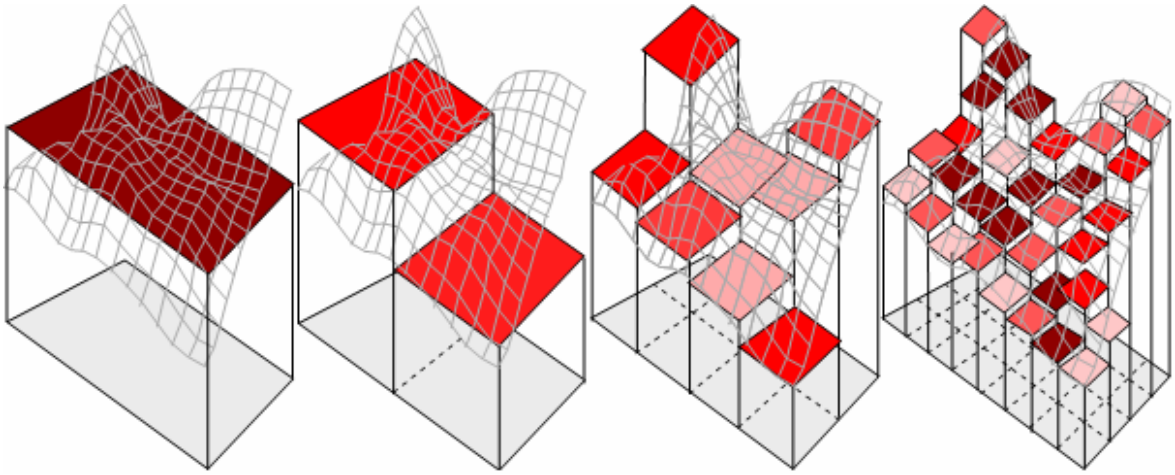

Figure A21 : Rugosité et taille des pixels

2.4. Régression multiple et géoréférencement

Un dernier exemple d'application des régressions multiples est consacré à la présentation des géoréférencements fondés sur un modèle polynomiale. Le principe d'un géoréférencement consiste à modifier les coordonnées d'une image ou d'un vecteur pour la rendre compatible avec un autre système de coordonnées, en l'occurrence celui d'une carte. Par exemple, le cadastre initialement levé par triangulation géodésique sur le terrain doit être rectifié pour être compatible avec les cartes à grande échelle de l'IGN. De même, une photographie aérienne et/ou une image de satellite doivent être géoréférencées pour

épouser les formes de la carte. Nous verrons plus loin que si les équations nécessaires à de telles transformations sont les mêmes pour les images ou les vecteurs, leur mise en œuvre posera plus de problèmes pour les images. À ce sujet, notons dès à présent qu'une correction géométrique et un géoréférencement sont deux notions différentes trop souvent confondues. La première a pour but de corriger la géométrie de l'image qui, à l'état brut, est très perturbée par la combinaison :

- des écarts d'attitude du satellite (lacet, roulis et tangage) ;
- des propriétés de l'orbite, elles-mêmes fonction du géoïde ;
- de la rotondité de la terre (déformation panoramique) ;
- de la vitesse de rotation de la terre (1669.79 $km.h^{-1}$ à l'équateur, soit 4.17 km parcourus pendant les 9 s que durent l'acquisition d'une image Spot, par exemple) ;
- de l'angle de prise de vue des pixels composants l'image ;
- des erreurs de parallaxe dues à la forme du relief (corrigées en intégrant les données d'un MNT : orthorectification) ;
- des erreurs liées à l'optique du capteur ;

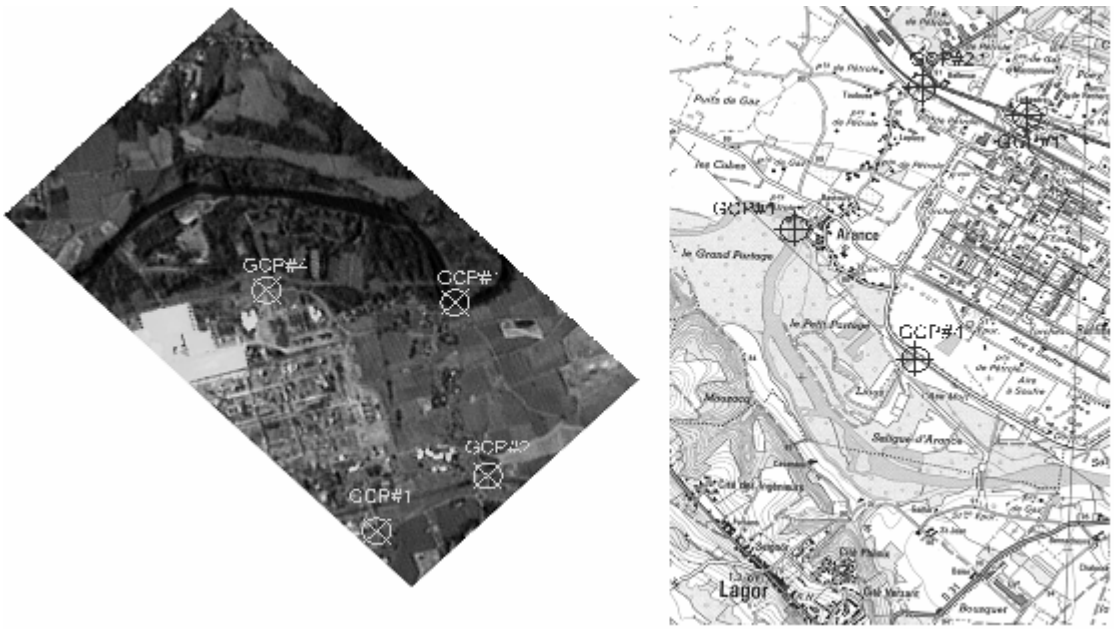

Figure A22 : Prise de points d'appui

Selon les cas – et les moyens financiers dont on dispose – tout ou partie de ces perturbations seront corrigées. Très souvent, certaines des corrections sont réalisées en même temps que le géoréférencement, d'où la confusion citée plus haut.

La mise en œuvre d'un géoréférencement consiste à relever les coordonnées de points identifiables sur les deux documents en question, la carte de référence (latitude et longitude) et l'image ou le vecteur (colonne et ligne) comme l'illustre la figure A22. Il est souhaitable de répartir au mieux les points sur l'ensemble des zones en commun dans les deux documents. On obtient la matrice de coordonnées suivante :.

id	Ligne	Colone	Latitude	Longitude	Résidus_Lat	Résidus_Lon	RMS error
GCP #1	2748.625	-421.625	436662.751	8767151.398	-6.998	-0.239	7.002
GCP #2	2801.625	-332.625	436790.594	8767339.312	-8.420	-0.150	8.422
GCP #3	2510.625	-161.625	436152.919	8767769.050	-22.920	1.525	22.971
GCP #4	2394.625	-33.625	435934.200	8768078.646	-17.428	5.964	18.420
GCP #5	2437.625	-127.625	436006.593	8767853.765	-19.532	0.876	19.552
GCP #6	2490.625	-281.625	436100.550	8767501.042	-17.309	-1.195	17.350
GCP #7	1187.625	-1215.625	433133.975	8765660.410	-0.072	2.923	2.924
GCP #8	551.625	-1159.625	431741.564	8765919.176	-7.176	17.222	18.658
GCP #9	635.625	-1255.625	431901.753	8765672.732	-9.546	7.334	12.038
GCP #10	511.625	-1635.625	431550.569	8764859.465	-16.749	8.520	18.792
GCP #11	107.625	-1111.625	430761.947	8766067.043	-14.420	3.250	14.781
GCP #12	131.625	-979.625	430835.880	8766381.260	-15.843	15.548	22.198
GCP #13	751.625	-2243.625	431975.686	8763473.215	-14.282	5.003	15.133
GCP #14	1655.625	-1987.625	434021.175	8763898.332	-1.955	9.200	9.405
GCP #15	2135.625	-2411.625	434969.986	8762900.232	-14.740	14.372	20.587
GCP #16	2380.78442	-2448.176	435509.350	8762791.650	-10.492	19.925	22.518
GCP #17	2291.625	-2469.625	435308.776	8762734.952	-10.655	9.084	14.001
GCP #18	1971.625	-2235.625	434646.551	8763304.117	-12.251	9.873	15.735
GCP #19	1487.625	-2311.625	433597.911	8763207.743	-2.305	7.352	7.705
GCP #20	2051.625	-2523.625	434776.387	8762658.228	-10.348	10.673	14.866
GCP #21	1683.625	-2295.625	434005.138	8763215.794	-13.175	8.586	15.726
GCP #22	1107.625	-2043.625	434007.829	8763203.846	564.471	-248.717	616.837
GCP #23	1379.625	-2571.625	433299.350	8762646.650	-10.446	4.120	11.229
GCP #24	1379.625	-3159.625	432844.367	8761405.308	-173.806	16.619	174.599
GCP #25	1975.625	-3203.625	434400.600	8761209.137	-54.321	24.044	59.405
GCP #26	2395.625	-2707.625	435466.895	8762209.801	-24.134	15.326	28.589
GCP #27	2487.625	-3153.625	435604.275	8761202.281	-18.037	11.148	21.204
GCP #28	2655.625	-3265.625	435953.229	8760923.118	-17.307	8.795	19.413
GCP #29	2583.625	-3171.625	435807.303	8761151.524	-19.806	13.018	23.701
Somme							43.923

Un géoréférencement consiste à définir les termes de deux polynômes d'ajustement de manière à estimer les latitudes et les longitudes observées à partir des coordonnées lignes et colonnes du document à modifier. Par exemple, pour un polynôme de degré 1 :

$$Lat' = a_1 Lig * b_1 Col + \varepsilon_1$$
$$Lon' = a_2 Lig * b_2 Col + \varepsilon_2$$

où *Lat'* et *Lon'* : la latitude et la longitude estimées ;
Lig et *Col* : les coordonnées en ligne et en colonne de l'image ou du vecteur;
a_n, b_n et ε_n : les termes du polynôme.

Notons qu'un minimum de points de contrôle est à prendre en fonction du degré du polynôme, ce seuil répond à la formule :

$$S = \frac{(t+1)(t+2)}{2}$$

où *t*, degré du polynôme.

Le parallèle avec les surfaces de tendances présentées plus haut est évident et l'on peut de la même manière élever le degré des polynômes pour obtenir un meilleur ajustement. Notons cependant que la fiabilité du modèle n'est plus estimée par le coefficient de détermination mais par un incateur quadratique moyen (*Root Meau Square* ou *RMS*) donnant les disatnces entre la position observée et celle estimée :

$$RMS_n = \sqrt{(lat'_n - lat_n)^2 + (lon'_n - lon_n)}^2$$

où *n* : numéro du point ;
lat' et *lon'* : latitude et longitude estimées ;
lat et *lon* : latitude et longitude observées.

Les lignes surlignées en couleur dans la matrice indiquent des points aberrants qu'il conviendrait d'éliminer de la collection initiale. Ces aberrations sont souvent dues à des erreurs de saisies ou à des mauvaises interprétations entre l'image et la carte ou bien encore à grande différence d'échelle – voire de qualité – entre l'image et la carte. Après vérification de la qualité du modèle celui-ci est appliqué à l'image pour obtenir une nouvelle image désormais correctement géoréférencée comme l'illustre la figure A23.

Figure A23 : Image corrigée et carte IGN

Le géoréférencement fondé sur un modèle polynomiale atteint rapidement ses limites puisqu'il n'intègre pas les altitudes sources d'importantes erreurs de parallaxe. Il est néanmoins facile à mettre en œuvre et efficace avec des données vectorielles ou des images de secteurs offrant peu de dénivelé.