Chapitre 1 : Corrélations et analyse des données

Méthode basée sur la corrélation

5.1-Généralités

La régression et la corrélation consistent en l'étude des liaisons existent entre deux ou plusieurs variables. En hydrologie, elles constituent l'outil mathématique le plus ancien et le plus largement utilisé, dont les buts sont multiples:

- Extension dans le temps des séries d'observations hydrologiques de courtes durées ;
- Prévision des grandeurs hydrologiques (écoulement à partir des conditions hydrométéorologiques observées : pluies, températures…..);
- Estimation des données manquantes dans les séries d'observations hydrologiques
- Etude de la dépendance entre les valeurs successives d'une série de données hydrologiques (série chronologiques).

On dit qu'il ya une corrélation entre deux variables observées, lorsque les variations des deux variables se produisent dans le même sens (corrélation positive), ou lorsque les variations sont dans le sens contraire (corrélation négative).

5.2 - Définitions

- **Régression**: c'est une méthode de recherche d'une relation exprimant le lieu entre une variable dépendante Y et une ou plusieurs variables dites indépendantes.
- **Corrélation :** c'est une méthode de recherche de la liaison qui existe entre deux variables aléatoires.

On peut calculer la corrélation existant entre n'importe quelles variables aléatoires. Des corrélations très élevées mais qui n'ont aucune signification sont très fréquentes, donc on n'entreprend une corrélation que lorsque la dépendance entre les variables peut être expliquée.

❖ **Diagramme de dispersion**

L'existence d'une corrélation entre deux variables peut être décelée graphiquement. Il s'agit de reporter les couples d'observations (x_i,y_i) sur un graphique en prenant pour abscisse la variable x, et pour ordonnée la variable y. Chaque point du graphique représente simultanément la valeur x_i, et la valeur y_i .Le graphique résultant constitue un nuage de points appelé : Diagramme de dispersion.

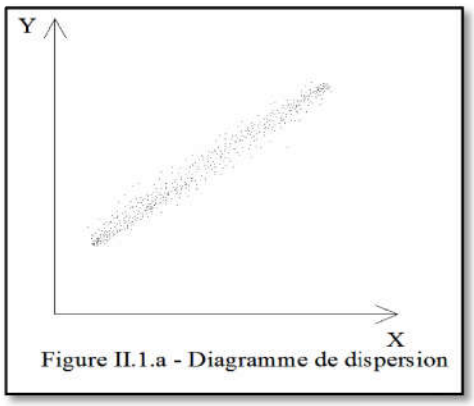

Figure II.1.a - Diagramme de dispersion

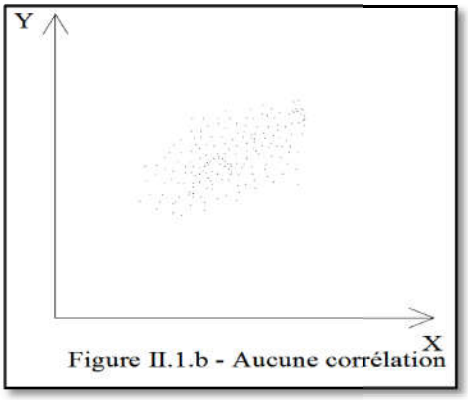

Figure II.1.b - Aucune corrélation

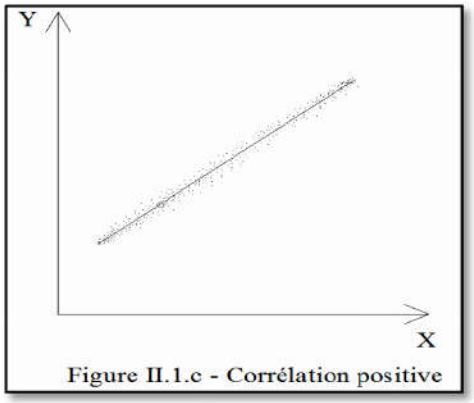

Figure II.1.c - Corrélation positive

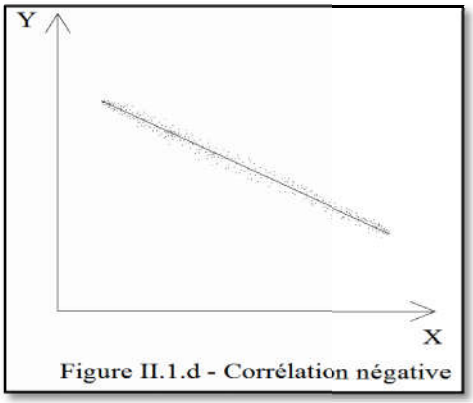

Figure II.1.d - Corrélation négative

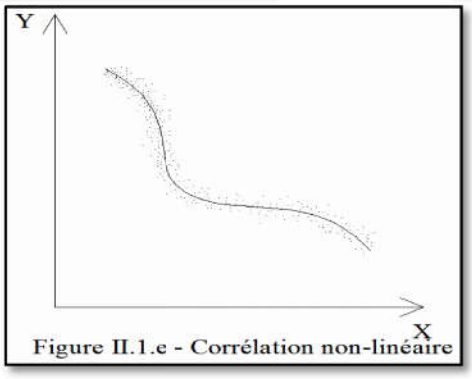

Figure II.1.e - Corrélation non-linéaire

schémas explicatifs

5.3 - Choix du modèle de régression

Lorsque le diagramme de dispersion est linéaire ou approximativement linéaire, on peut s'efforcer de rechercher l'équation de la droite qui s'y ajuste le mieux. Cette droite de régression de régression de Y en X est généralement déterminée par la méthode des moindres carrées.

Dans la pratique, on s'efforce toujours de trouver une régression linéaire même s'il faut faire une transformation dans la relation fonctionnelle. Les différents modèles existant sont:

Le modèle linéaire représenté par l'équation de la droite : $Y=A+BX$

Les modèles curvilinéaires, à savoir :

Le modèle puissance $Y=AX^B$

Le modèle exponentiel $Y=Ae^{BX}$

Le modèle parabolique $Y=A+BX+CX^2$

5.5-Régression linéaire simple

5.5 .1 - Coefficient de corrélation

C'est l'indice qui mesure l'intensité de la liaison linéaire entre deux variables, qui est un nombre sans dimension, il est donné par :

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

En raison de la symétrie de sa définition, le coefficient de corrélation mesure aussi bien l'intensité de la liaison entre y et x qu'entre x et y.

Le coefficient de corrélation est indépendant des unités de mesure de x et de y.

La valeur du coefficient de corrélation peut varier entre -1, (corrélation négative et parfaite) et +1 (corrélation positive et parfaite). Plus les points sont étroitement alignés selon

une droite, plus la valeur du coefficient de corrélation sera élevée et s'approchant de +1 ou -1 selon le cas).

5.5.2 - Droite de régression-méthode des moindres carrés

La droite de régression c'est la droite qui s'ajuste le mieux aux observations, elle constitue un outil de prévision. On pourra estimer ou prévoir, à l'aide de cette équation, les valeurs d'une variable à partir des valeurs prises par l'autre variable.

Pour la régression linéaire, la droite de régression de Yen X est généralement déterminée par la méthode des moindres carrés, qui consiste à minimiser la somme des carrés des écarts entre les points observés et les points correspondants sur la droite.

Soit un échantillon de n couples d'observations (x_i,y_i) et soit l'équation de la droite :

$$\hat{y}= b_0 + b_1x_i$$

Où :

b_0 : Ordonnée à l'origine;

b_1 : Pente de la droite ;

$\hat{y}$: Représente la valeur estimée de la variable dépendante pour une valeur particulière x_i de la variable explicative (indépendante).

Soit ei l'écart verticale entre la valeur observée y_i et l'estimation $\hat{y}$ obtenue par la droite de régression pour $X=x_i$.

$ei= y_i- \hat{y} = y_i- b_0 - b_1x_i,$ pour i=1,….n.

La Somme des carrés de ces écarts pour l'ensemble des points est égale à:

$S= e1^2 + e2^2 +\ldots\ldots\ldots\ en^2 = \sum_{i=1}^{n}(yi \quad \hat{y})^2 = \sum_{i=1}^{n}(yi \quad b0 \quad b1xi)^2$

La méthode des moindres carrés permet de déterminer les expressions b_0 et b_1 de telle sorte que la somme S soit minimale. La droite obtenue est dite droite des moindres carrés, ou droite de régression. On trouve:

$y = ax + b$ $\qquad a = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ $\qquad b = \bar{y} - a\bar{x}$

Où :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad ; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad ; \quad S_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \quad ; \quad S_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}}$$