Exercice 5

HOMOGENEISATION DES DONNEES

Soient 2 stations pluviométriques "Menaceur" et "Lazabane" situées à quelques kilomètres l'une de l'autre dans le bassin versant du côtier Algérois. Ces stations ayant fonctionné respectivement sur des périodes de 20 ans (N) et de 14 ans (K) comme le montre le tableau 5.1. En supposant que la série pluviométrique des précipitations annuelles de la station de Menaceur est la station de référence (X) et que l'erreur recherchée se trouve au niveau de la série pluviométrique de Lazabane, série à étudier (Y), on demande de :

1- Vérifier 'homogénéité de la série de la station de référence (X) en appliquant le test de la médiane ;
2- Détecter l'erreur systématique de la station étudiée et faire la correction par la méthode des doubles masses s'il y a erreurs ;
3- Projeter sur un diagramme de dispersion les séries avant et après correction
4- Vérifier l'homogénéité de la série Y après correction en appliquant le test de Wilcoxon ;
5- Donner la droite de régression de Y en X ;
6- Calculer le gain obtenu pour l'extension ;
7- Faire l'extension de la série courte à la série longue
8- A partir de la méthode d'estimation, calculer les caractéristiques empiriques de la série étendue.

Données

Tableau 5.1. Données pluviométriques

Année	Station Menaceur	Station Lazabane	Année	Station Menaceur	Station Lazabane
1980-81	390	**???**	1990-91	317	251
1981-82	520	**???**	1991-92	554	462
1982-83	470	**???**	1992-93	778	689
1983-84	708	628	1993-94	408	356
1984-85	565	469	1994-95	520	220
1985-86	609	400	1995-96	646	301
1986-87	582	495	1996-97	762	305
1987-88	843	688	1997-98	430	**???**
1988-89	640	480	1998-99	594	**???**
1989-90	619	587	1999-00	707	**???**

Corrigé :

L'homogénéisation des données consistent à identifier les séries pluviométriques et à vérifier s'il n'y a pas d'erreurs systématiques qu'il convient de rechercher et de corriger s'il y a lieu.

Pour la fiabilité de l'information, il convient de tester la série de référence utilisée pour d'autres séries.

1. Test de la médiane : série de référence ou de base

Le test da la médiane (test de Mood) étant réalisé, l'homogénéité étant vérifiée, ce test permettra de voir si la série à étudier est homogène ou pas, c'est-à-dire si elle appartient à la même population que la série de référence.

Soit un échantillon x_1, x_2, x_3,……x_n ; déterminons sa médiane **m** après avoir classé l'échantillon par ordre croissant.

La médiane **m** est une constante de telle sorte que 50% des x_i lui soient inférieures et 50% des x_i lui soient supérieures.

Remplaçons donc la série des valeurs non classées par une suite de signe :

+pour les $x_i > \mathbf{m}$

-pour les $x_i < \mathbf{m}$

Calculons les quantités **Ns** et **Ts**, avec :

Ns : nombre total de séries de + ou de - dans la série initiale ;

Ts : taille de la plus grande série de + ou de - au-dessus de la médiane dans la série initiale.

Ns suit approximativement une loi normale de moyenne $\frac{N+2}{2}$ et de variance $\frac{1}{4}(N-1)$ et T suit une loi binomiale. Ceci a permis d'établir que pour un seuil de signification compris entre 91% et 95%, les conditions du test sont les suivantes :

$$Ns > \frac{1}{2}\left(N+1 - u_{1-\frac{\alpha}{2}}\sqrt{N+1}\right) \qquad (5.1)$$

$$Ts < 3.3\,(\log_{10} N + 1) \qquad (5.2)$$

Si les conditions du test sont vérifiées, on conclut que la série à étudier est homogène au seuil de signification $1 - \alpha$.

La médiane **m** déterminée sur la série Y classée par ordre croissant est :

m = 588 mm

L'application du test nécessite la vérification des conditions Ns et Ts (Tableau 5.1).

Tableau 5.1. Test de la médiane (série de référence)

N°	P_{an} (mm)	
1	390	-
2	520	-
3	470	-
4	708	+
5	565	-
6	609	+
7	**582**	-
8	843	+
9	640	+
10	619	+
11	317	-
12	554	-
13	778	+
14	408	-
15	520	-
16	646	+
17	762	+
18	430	-
19	594	-
20	707	-

Médiane = m = 588 mm

N=20

$u_{1-\alpha/2}$ = 1.96 (variable de Gauss, lu sur la table de Gauss pour un seuil de signification 1-α =95%.

$$Ns > \frac{1}{2}\left(N+1 - u_{1-\frac{\alpha}{2}} \sqrt{N+1} \right) = 6.01$$

Ns = 12 (déterminé sur la série initiale)

$Ts < 3.3 (\log_{10} N + 1) = 7{,}59$

Ts = 5 (déterminé sur la série initiale, maximum de + ou de - au-dessus de la médiane).

Conclusion :

Pour Ns : On a : 12 > 6.01

Pour Ts : On a : 5<7,59

La série de référence est homogène.

L'homogénéité de la série de référence étant vérifiée, cette série servira de base pour la détection d'erreurs systématiques dans la série à étudier.

Cependant, les stations pluviométriques à partir desquelles les séries sont considérées doivent appartenir aux mêmes conditions climatiques.

Il est important d'identifier la station de base ou de référence pour pouvoir détecter et corriger les erreurs de la station à étudier.

2. Méthode des doubles masses

Par moment, il est difficile d'identifier une station de référence. De ce fait, il faut répertorier sur un tableau toutes les stations appartenant au bassin versant considéré (Tableau 5.2), en précisant la période d'observations, le nombre d'années d'observation (N) et les coordonnées géographiques de chaque station (X, Y) et l'altitude (Z) et d'en créer une station fictive F qui sera considérée comme station de référence (Tableau 5.3). La valeur des précipitations sera la moyenne des pluies aux stations disponibles Son homogénéité doit être vérifiée par un test statistique (test de la médiane ou test de Wilcoxon par exemple).

Tableau 5.2. Identification des stations sur le bassin versant

Station	1971/72…….1987/88……..1999/00	N	X Km	Y km	Z m
A B C D	---------- ---------- ---------- ----------				

Tableau 5.3. Station fictive

Station / Année	A	B	C	D	Station fictive F
1971/72	X_{Ao}	X_{Bo}	X_{Co}	X_{Do}	$X_{Fo} = \frac{X_{A0} + X_{B0} + ... +}{n}$
1972/73	X_{A1}	X_{B1}	X_{C1}	X_{D1}	X_{F1}
.					.
.					.
.					.
N	X_{AN}	X_{BN}	X_{CN}	X_{DN}	X_{FN}

Si nous revenons à notre exercice, la station de référence est celle de Menaceur **(X).** Elle a été vérifiée par le test de la médiane. La station à étudier est celle de Lazabane **(Y)** qu'il convient de vérifier, de corriger en cas d'erreurs et d'étendre pour son utilisation future.
La méthode des doubles masses est considérée.

Les valeurs correspondantes à la même période d'observations sont reportées en coordonnées rectangulaires, obtenant une courbe de double cumul.

Si les données de la station contrôlée sont homogènes par rapport à celles de la station de base, la courbe des doubles cumuls avoisine une droite (Fig. 5.1). Si elle possède une cassure à partir d'un point *M*, les observations à partir de ce point sont hétérogènes.

Dans le cas où l'hétérogénéité est détectée, la correction s'effectue par modification de la pente de la droite de double cumul des données antérieures ou postérieures à la date de la cassure. Seul le but visé par l'étude en cours peut indiquer quelle partie de la série est à corriger.

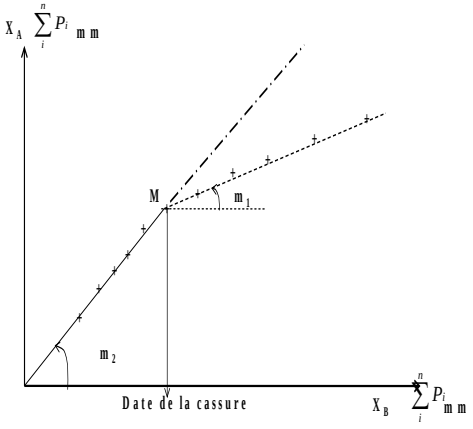

Fig. 5.1. Méthode des doubles masses

On corrige les données observées en multipliant le rapport de pente $\frac{m_1}{m_2}$ ou $\frac{m_2}{m_1}$ par la valeur erronée dans la série observée respectivement selon que l'on soit après la cassure ou avant la cassure.

Procédé :

Le tableau 5.4 représente les valeurs initiales et cumulées des précipitations annuelles aux 2 stations pluviométriques.

La méthode de la double masse appliquée aux cumuls annuels des 2 stations a permis de confirmer l'hétérogénéité de la série des pluies annuelles de la station Y comme le montre la figure 5.1.

Au vue de la figure 5.2, la station Y présente une hétérogénéité qu'il convient de corriger.

Le changement ou la cassure de la pente correspond à l'année 1993/94. A partir de cette année, les 3 autres années qui suivent sont erronées et doivent être rectifiées.

Les pentes m_1 et m_2 correspondant respectivement aux 1er et 2ème segments de droite sont calculées :

$m_1 = 0.82$

$m_2 = 0.43$

Le choix de la période à corriger est un peu arbitraire si on ne dispose pas des originaux. Cependant, soit on peut corriger la période la plus courte soit corriger les données antérieures à la rupture.

Dans cet exemple, le choix a porté sur la période après la date de la cassure, en corrigeant les 3 dernières années de la station Y par un coefficient multiplicatif (rapport $m_1/m_2 = 1,91$) (Tableau 5.4).

Tableau 5.4. Cumul annuel des 2 stations

Année	P_{an} (**X**) (Station de base) mm	P_{an} (**Y**) (Station à étudier) mm	Cumul **X** mm	Cumul **Y** mm
1983-84	708	628	708	628
1984-85	565	469	1273	1097
1985-86	609	400	1882	1497
1986-87	582	495	2464	1992
1987-88	843	688	3307	2680
1988-89	640	480	3947	3160
1989-90	619	587	4566	3747
1990-91	317	251	4883	3998
1991-92	554	462	5437	4460
1992-93	778	689	6215	5149
1993-94	408	356	6623	5505
1994-95	520	**220**	7143	5725
1995-96	646	**301**	7789	6026
1996-97	762	**305**	8551	6331

La représentation graphique des cumuls est donnée en figure 5.2.

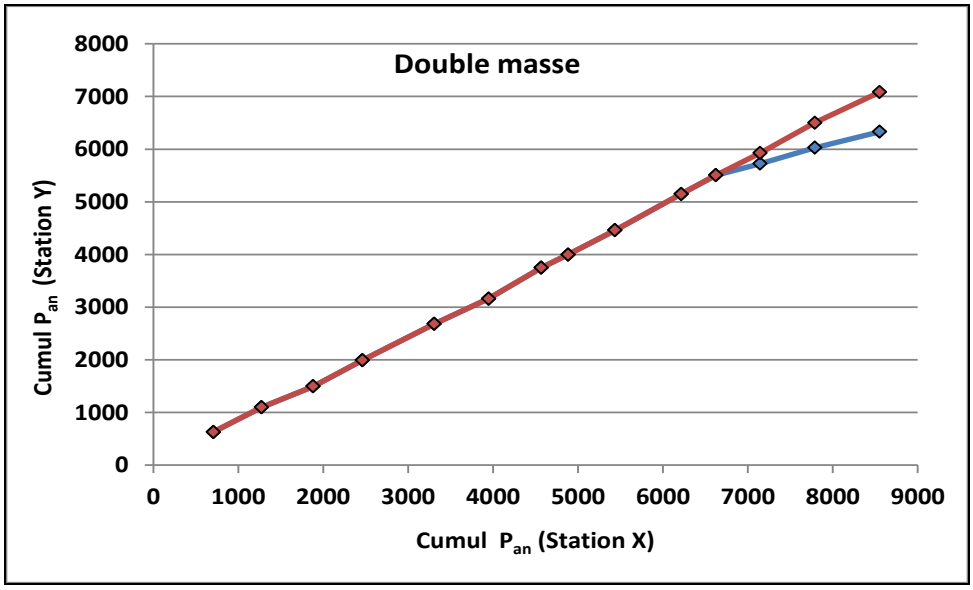

Fig.5.2. Double cumul

Tableau 5.5. Valeurs annuelles initiales et corrigées (Station Y)

Année	P_{an} (Y initial) mm	P_{an} (Y corrigée) mm
1994-95	220	420
1995-96	301	575
1996-97	305	583

3. Diagramme de dispersion (K années)

Les couples (X,Y) sont projetés sur un diagramme de dispersion pour la même période d'observation (Fig.5.3). Il est à noter que la projection des pluies annuelles observées sur un diagramme de dispersion montre d'une part, l'existence d'une dispersion des points aboutissant à une faible corrélation entre les 2 stations. Le coefficient de détermination $R^2 = 0.41$ (Fig.5.3) confirme notre hypothèse. Trois valeurs ont déstabilisent la corrélation (220, 301 et 305) correspondant respectivement aux années 1994/95, 1995/96 et 1996/97.

Après correction des valeurs erronées de la station Y, la relation est plus nette, le coefficient de détermination (carré du coefficient de corrélation) est significatif ($R^2 = r^2$). Il est passé de 0,41 (Fig.5.3) à 0.87 (Fig.5.4).

La droite de régression régissant la relation Y en X est :

$$P_{an,Y} = 0.83\ (P_{an},x) - 6.14. \qquad (5.3)$$

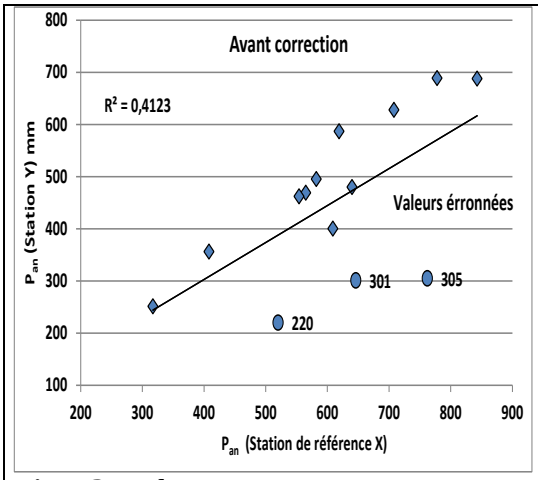

Fig.5.3. Relation Y-X avant correction

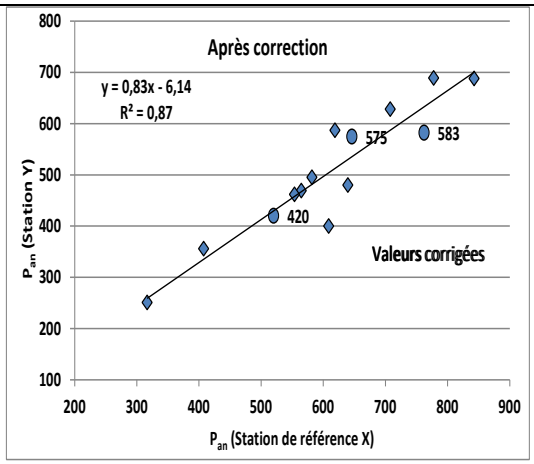

Fig.5.4. Relation Y-X après correction

Cette droite va permettre de faire l'extension de la série Y, c'est-à-dire combler les lacunes.

4. Test de Wilcoxon ou Test des rangs

Avant de faire l'extension, il convient de tester si la série corrigée appartient à la même population que la série de référence. Le test de Wilcoxon est le plus puissant des tests non paramétriques.

Rappel : Soient 2 variables aléatoires Y et *X*, représentant respectivement 2 séries de précipitations annuelles de taille N_1 et N_2.

Y étant la série à étudier et X étant la série de base avec $N_2 > N_1$.

Si l'échantillon *Y* est issu de la même population que l'échantillon X, l'échantillon nouveau Y U X est également issu de la même population. De ce fait, on classe les éléments de ce nouvel échantillon Y U X par ordre croissant et on attribue à chacune des valeurs le rang qu'elle occupe dans cette nouvelle série. (Si une valeur se répète plusieurs fois, il faut lui associer le rang moyen qu'elle détermine).

On calcule les quantités W_Y et W_X :

W_Y représente la somme des rangs de Y et c'est celle qui nous intéresse et est égale à :

$$W_Y = \sum_{i=1}^{n} \text{rang}_Y \quad = 1+ 3 + 4 + \ldots\ldots + 13 + 17 + \ldots + n \qquad (5.4)$$

$$W_X = \sum_{i=1}^{n-1} \text{rang}_X \; = 2 + 5 + \ldots + 12 + 14 + 15 + 16 + \ldots + n\text{-}1 \qquad (5.5)$$

L'hypothèse nulle est vérifiée si :

$$W_{min} < W_Y < W_{max} \qquad (5.6)$$

Avec :

$$W_{min} = \frac{(N_1 + N_2 + 1)\, N_1 - 1}{2} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}} \qquad (5.7)$$

Et,

$$W_{max} = (N_1 + N_2 + 1)N_1 - W_{min} \qquad (5.8)$$

$u_{1-\frac{\alpha}{2}}$ représente la valeur de la variable centrée réduite de Gauss correspondant à une probabilité de $1 - \frac{\alpha}{2}$.

L'hypothèse d'homogénéité est acceptée si l'égalité suivante est vérifiée :

$$W_{min} < W_Y < W_{max}$$

Le procédé de calcul est présenté dans le tableau 5.6.

La condition d'appartenance à la même population, la série corrigée est à considérer.

5. Droite de régression

Après correction des valeurs erronées de la station Y, la relation est plus nette, le coefficient de détermination (carré du coefficient de corrélation) est significatif ($R^2 = r^2$). Il est passé de 0,41 (Fig.5.5) à 0.87 (Fig.5.6).

La droite de régression régissant la relation Y en X est :

$$P_{an,\mathbf{Y}} = 0.83\ (P_{an,\mathbf{X}}) - 6.14. \qquad (5.9)$$

Tableau 5.6. Valeurs initiales

N	Station Menaceur	Station Lazabane
1	390	
2	520	
3	470	
4	708	***628***
5	565	***469***
6	609	***400***
7	582	***495***
8	843	***688***
9	640	***480***
10	619	***587***
11	317	***251***
12	554	***462***
13	778	***689***
14	408	***356***
15	520	***420***
16	646	***575***
17	762	***583***
18	430	
19	594	
20	707	

$$W_Y = \sum_{i=1}^{n} \text{rang}_Y = 392$$

$$W_X = \sum_{i=1}^{n-1} \text{rang}_X = 203$$

$W_{min} = 293$

$W_{max} = 406$

Condition du test vérifiée

$293 < W_Y < 406$

Les 2 séries appartiennent à la même population et sont homogènes

Tableau 5.7. Test de Wilcoxon

X U Y	Rang	Somme Rang Y	Somme Rang X
390	1	1	
251	***2***		2
317	3	4	
356	***4***		6
400	***5***		11
408	6	10	
420	***7***		18
430	8	18	
462	***9***		27
469	***10***		37
470	11	29	
480	***12***		49
495	***13***		62
520	14	43	
520	15	58	
554	16	74	
565	17	91	
575	***18***		80
582	19	110	
583	***20***		100
587	***21***		121
594	22	132	
609	23	155	
619	24	179	
628	***25***		146
640	26	205	
646	27	232	
688	***28***		174
689	***29***		**203**
707	30	262	
708	31	293	
762	32	325	
778	33	358	
843	34	**392**	

6. Calcul du gain

Avant de combler les lacunes, il convient de chercher la taille de la nouvelle série, c'est-à-dire, il faut calculer le gain. Les séries étant de tailles différentes, il est difficile d'étendre la série courte à la série longue sans préalablement connaître jusqu'à combien de valeurs peut-on étendre la série courte.

Rappel : Le bénéfice de l'extension de la série **Y** à l'aide de la série **X** *pour* la connaissance de la série *Y* est d'autant plus grand que le coefficient de corrélation $_{k}\mathbf{r}_{xy}$ est élevé.
Ce bénéfice a été traduit par R.Véron en efficacité relative **E**.

$$E = 1 + (1 - \frac{K}{N}) \left[\frac{1-(k-2)\,r^2}{k-3}\right] \qquad (5.10)$$

$r = {}_{k}r_{xy}$: coefficient de corrélation calculé sur K années

E : Efficacité relative de $\overline{y_k}$ et de $\overline{\hat{y}}$ définie par le rapport de la variance de $\overline{\hat{y}}$ à celle de $\overline{y_k}$.

Ce bénéfice est traduit, en utilisant E sous la forme d'un gain réel d'information que l'on exprime à l'aide du *nombre d'années efficaces* ou *fictives* N'à laquelle correspond l'échantillon Y étendu.

N' varie de K à N (gain maximum, liaison fonctionnelle entre X et Y et r =1).

$N' = \frac{K}{E}$ avec $K \geq 3$; On obtient une nouvelle série qui au total fait bien N mais à laquelle, on ne peut attribuer la même confiance que N'années obervées.

Pour notre exercice :

K = 14
N=20
$R^2 = 0.87$ d'où r = 0,93
E = 0, 745
N' = 18,85 = 19

Ceci représente une série de vraies valeurs observées dans laquelle on pourrait avoir la même confiance que dans les valeurs (14 observations et 6 valeurs reconstituées).

Lors du calcul de l'intervalle de confiance au lieu de considérer 20 valeurs on considèrera N' valeurs soit 19.

7. Extension de la série Y

Le choix restera aléatoire, de préférence, il faut combler les lacunes d'abord par les années plus récentes en utilisant l'équation de la droite de régression calculée précédemment.

La relation fonctionnelle régissant la régression linéaire de Y en X est appliquée.

$$P_{an,Y} = 0{,}83\ (P_{an,X}) - 6{,}14.$$

Les résultats sont présentés le tableau 5.8.

Tableau 5.8. Extension de la série Y après calcul du gain

	Nbre	Année	Station Menaceur P_{an} (X) mm	Station Lazabane P_{an} (Y) mm	Observations
	1	1980/81	390	**318**	Valeurs étendues
	2	1981/82	520	**425**	
	3	1982/83	470	**384**	
	4	1983/84	708	628	
	5	1984/85	565	469	
	6	1985/86	609	400	
	7	1986/87	582	495	
	8	1987/88	843	688	
	9	1988/89	640	480	
K	10	1989/90	619	587	
	11	1990/91	317	251	
	12	1991/92	554	462	
	13	1992/93	778	689	
	14	1993/94	408	356	
	15	1994/95	520	**420**	Valeurs corrigées
	16	1995/96	646	**575**	
	17	1996/97	762	**583**	
	18	1997/98	430	**351**	Valeurs étendues
	19	1998-99	594	**487**	
	20	1999/2000	707	**581**	

8. Méthode d'estimation sur la série étendue

Une autre méthode peut être utilisée quand les séries sont très longues, c'est la méthode de l'estimation.

Rappel :

Soient 2 séries hydrologiques de hauteurs annuelles de précipitations :

X : série longue constituant la série de base à **N** observations

Y : série courte ou série à étudier à **K** observations (K < N)

Les estimations des paramètres statistiques et des valeurs annuelles d'une série X de N valeurs observées peuvent être estimées sans pour autant passer par la méthode classique.

Moyenne estimée par l'extension $\hat{\bar{y}}$

$$\hat{\bar{y}} = \bar{y}_k + {}_k r_{xy} \frac{{}_k S_y}{{}_k S_x} (\bar{x}_n - \bar{x}_k) \qquad (5.11)$$

$\hat{\bar{y}}$: Estimation de la moyenne de la série des *y* (dont l'espérance mathématique est $\bar{y}$ (toujours inconnue) à partir de la 1ère estimation de $\bar{y}_k$, des autres paramètres statistiques des échantillons des valeurs observées et du coefficient de corrélation entre x et *y*.

Variance estimée par l'extension $\hat{\sigma}^2$

$$\hat{\sigma}^2 = {}_K S_y^2 + {}_K r^2{}_{xy} \frac{{}_k S_y^2}{{}_k S_x^2} ({}_N S_x^2 - {}_K S_x^2) \qquad (5.12)$$

$\hat{\sigma}^2$: estimation de la variance de Y (dont l'espérance mathématique est σ_y^2) à partir de la 1ère estimation de ${}_K S^2{}_y$ et des estimations de la variance de X et du coefficient de corrélation entre X et Y.

${}_n S^2{}_x$: variance de X à partir de l'échantillon de N valeurs.

Coefficient de corrélation estimé

$$\hat{r} = {}_k r_{xy} \frac{{}_K S_y}{{}_K S_x} \frac{{}_N S_x}{\hat{\sigma}_y} \qquad (5.13)$$

$\hat{r}$: Estimation du coefficient de corrélation entre X et Y (dont l'espérance mathématique est r) à partir de la 1ère estimation ${}_K r_{xy}$; des écarts types des échantillons de valeurs observées et de l'estimation $\hat{\sigma}_y$ définie ci-dessus.

NB : En pratique, comme il est difficile de tester la signification relative de $\hat{r}$ et de ${}_K r_{xy}$, on conserve le coefficient de corrélation expérimental ${}_K r_{xy}$..

Valeur estimée par l'extension

$$Y_x = \bar{y}_x + {}_K r_{xy} \frac{{}_K S_y}{{}_K S_x} (x - \bar{x}_k) \qquad (5.14)$$

Avec :

y_x : moyenne conditionnelle de *Y* liée à *X*

$\bar{x}_k$ et $\bar{y}_k$: Moyenne arithmétique de X et de Y calculée respectivement à partir de k valeurs observées simultanément.

${}_K r_{xy}$: Coefficient de corrélation entre *X* et *Y* sur k années d'observations communes

${}_K S_x$ et ${}_K S_y$: Ecart type de X et de Y sur k années observées communes

Les caractéristiques empiriques des 2 séries sont résumées dans le tableau 5.9.

Moyenne de la série Y estimée par l'extension pour N Valeurs

$$\hat{\bar{Y}} = {}_K \bar{Y} + {}_k r_{xy} \frac{{}_k S_Y}{{}_k S_X} ({}_N \bar{X} - {}_K \bar{X}) = 482 \text{ mm}$$

Variance de la série Y estimée par l'extension pour N valeurs

$$\hat{\sigma}^2 = {}_K S_y^2 + {}_K r^2_{xy} \frac{{}_K S_y^2}{{}_K S_x^2} ({}_N S_x^2 - {}_K S_x^2) = 14932$$

Ecart type estimé $= \sqrt{\hat{\sigma}^2} = 122$ mm

Coefficient de corrélation estimée par l'extension pour K valeurs

$${}_K \hat{r} = {}_K r_{xy} \frac{{}_K S_y}{{}_K S_x} \frac{{}_N S_x}{\hat{\sigma}_y} = 0{,}95$$

Coefficient de corrélation estimée par l'extension pour N valeurs

$${}_N \hat{r} = {}_K r_{xy} \frac{{}_K S_y}{{}_K S_x} \frac{{}_N S_x}{\hat{\sigma}_y} = 0{,}95$$

Valeur estimée par l'extension pour N valeurs

$$ {}_{N}Y_{x} = \bar{y}_{x} + {}_{K}r_{xy} \frac{{}_{K}S_{y}}{{}_{K}S_{x}} (x - \bar{x}_{k}) $$

Exemple : Année 1998/97 : X = 430mm d'où $Y_{1998/97}$ = 351 mm

Les résultats sont résumés dans le tableau 5.9.

Tableau 5.9. Caractéristiques des séries initiale et étendue

Station	Station X	Station Y	
		Valeurs	
		initiale	estimée
Taille	20	14	20
Somme	11662	7083	9627
Moyenne	${}_{N}\overline{X}$ = 583 ${}_{K}\overline{X}$ = 611	${}_{K}\overline{Y}$ = 506	${}_{N}\overline{Y}$ = 482
Ecart Type	${}_{N}S_{X}$ = 138 ${}_{K}S_{X}$ = 141	${}_{K}S_{Y}$ = 127	${}_{N}S_{Y}$ = 122
Variance	${}_{N}S_{X}^{2}$ =19046 ${}_{K}S_{X}^{2}$ = 19860	${}_{K}S_{y}^{2}$ = 16069	${}_{N}S_{y}^{2}$ = 14932
${}_{K}r_{xy}$	0,93		
${}_{N}r_{xy}$	0,95		

NB : Pour cette méthode, nous n'avons pas tenu compte du gain. Nous avons considéré la série étudiée et étendue comme étant égale à 20 et non à 19 valeurs.

&&&&&