

# Analyse Statistique de données altimétriques type SWOT et bases de données auxiliaires

Manon Leclère, stagiaire 3e année MIC, INSA de Toulouse

16 juin 2019 - 11 septembre 2019

Sous le tutorat de M. Kevin Larnier et sous la responsabilité de M. Jerome Monnier

## Introduction

Dans le cadre de la mission spatiale SWOT à venir, les chercheurs de l'IMT et du CS ont élaboré un modèle physique pour inférer le débit de rivières mondiales, à l'aide de données altimétriques de type SWOT. Avant d'être effectif, le modèle doit passer par une phase d'apprentissage à l'aide d'une base de données altimétriques réelles de type SWOT, mesurées de rivières d'Amérique du Nord. Pour pouvoir effectuer cet apprentissage, il est important de connaître les corrélations entre les variables d'intérêt de cette base de données. Cependant, on ne peut pas l'exploiter telle quelle : il est nécessaire d'explorer au préalable les données, de déceler les problèmes et de trouver des conditions dans lesquelles elles sont utilisables. On s'intéresse ici à cette problématique. On effectuera donc l'étude de « HYDRoSWOT\_100m », base de données de type SWOT à explorer, constituée de mesures réelles sur des rivières d'Amérique du Nord de plus de 100m de largeur ; en parallèle de deux bases de données mondiales « PEPSI1 » et « PEPSI2 », qui serviront de référence pour validation et dont les mesures proviennent de sorties de modèles numériques.

Dans un premier temps, on décrira simplement les trois jeux de données et les variables étudiées. Dans un second temps, on effectuera des tests statistiques entre les échantillons géographiquement comparables de HYDRoSWOT\_100m et PEPSI1/PEPSI2 pour vérifier la cohérence et l'exploitabilité des données. Dans un troisième temps, on analysera les corrélations entre les variables des jeux de données, on effectuera plusieurs Analyses en Composantes Principales ainsi qu'une Analyse par Classification Ascendante Hiérarchique (clustering). Enfin, on comparera brièvement les corrélations des rivières américaines avec celles des rivières européennes et asiatiques de PEPSI.

## 1 Description des jeux de données HYDRoSWOT\_100m et PEPSI

### 1.1 Géolocalisation des stations de mesures

Les bases de données sont, par définition, constituées d'observations de différentes variables d'intérêt sur les individus d'une population définie. Dans le cas de HYDRoSWOT\_100m, les observations sont faites sur différents sites de mesures répartis sur des rivières américaines de plus de 100 mètres de largeur. Dans le cas de PEPSI1 et PEPSI2, les observations sont des sorties numériques générées pour un point d'une rivière à des instants différents. On considère que les individus sont les stations de mesures des rivières et que la population est constituée de l'ensemble de ces stations, et par abus de langage, des rivières. Les bases de données PEPSI concernent des rivières mondiales, cependant dans la suite sauf précision du contraire, on décrira seulement les rivières d'Amérique du Nord.

Voici en Figure 1 et Figure 2 deux cartes de l'Amérique du Nord générées par QGIS. La Figure 1 est une carte satellite ; la Figure 2 est une représentation graphique du chevelu des rivières d'Amérique du Nord. On situe sur ces cartes les différentes stations de mesures pour HYDRoSWOT\_100m, en bleu ; les stations de sorties de PEPSI1 en vert et PEPSI2 en rouge.

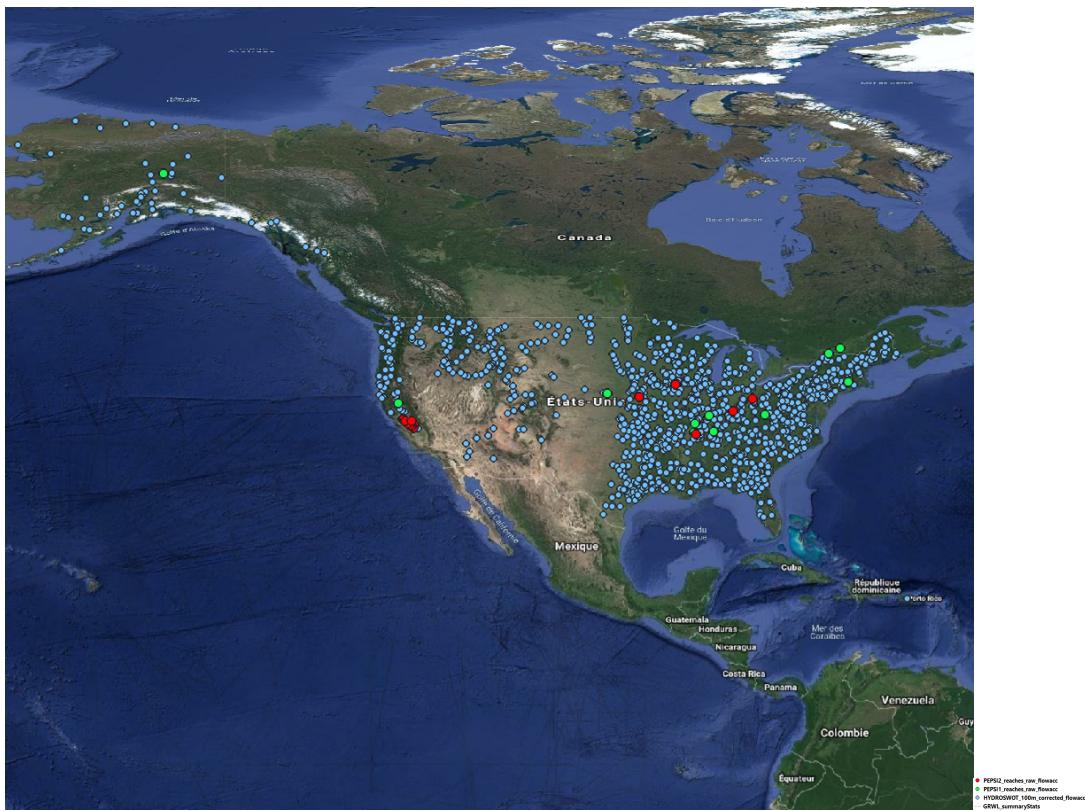


FIGURE 1 – QGIS - HYDROSWOT\_100m, PEPSI1 & PEPSI2 Geolocalisation / Satellite map

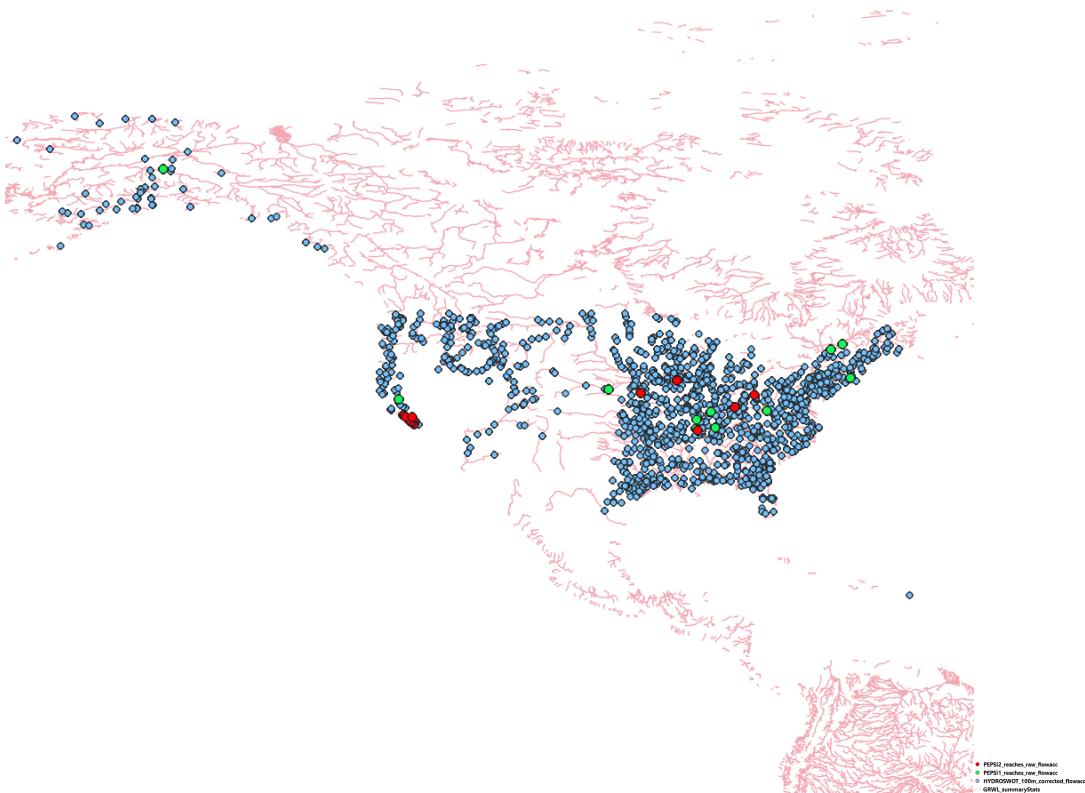


FIGURE 2 – QGIS - HYDROSWOT\_100m, PEPSI1 & PEPSI2 Geolocalisation / River map

On peut d'ores et déjà constater que géographiquement, les données ne sont pas équilibrées entre HYDROSWOT\_100m et les PEPSI. La géolocalisation ainsi que la quantité d'observations par station et par rivière vont être les principales sources de contraintes à surmonter pour obtenir des échantillons de données comparables et analyser correctement les jeux de données.

## 1.2 Variables étudiées

- Pour chaque station, chaque individu, on peut observer la valeur des variables suivantes :
- Les variables les plus importantes pour le modèle physique d'inférence du débit des rivières :
    - « Q » le débit de la rivière ( $m^3/s$ )
    - « flowacc » l'aire draînée, ( $m^2$ )
    - « dA » les aires ajoutées ( $m^2$ ),
    - « W » la largeur (m),
    - « stage » l'élévation (m),
    - « dH » les variations d'élévations (m),
    - « S » la pente, variable observée uniquement chez PEPSI1 et PEPSI2 (cm/km),
  - Des variables importantes pour l'hydrologie, mais plutôt utilisées ici à titre exploratoire :
    - Les variables précisant la constitution de la terre du lit de la rivière (%) :
      - “clay” le taux d'argile ,
      - “sand” le taux de sable,
      - “silt” le taux de limon,
    - Les 12 variables faisant un portrait du sol aux alentours de la station, les “Land Cover”. Elles sont divisées en 12 classes (%) :
      - le taux de présence de différents types de végétation : les arbres de LC1 à LC4, les arbustes LC5, les herbacées LC6,
      - le taux de terres cultivées aux alentours, LC7
      - le taux de terres souvent immergées, LC8
      - le taux de présence de construction humaine, d'urbanisme, LC9
      - le taux de présence de neige/glace, LC10
      - le taux de terres stériles/arides autour, LC11
      - le taux de présence d'eau, LC12

Bien entendu, il reste les variables d'identification des stations, ainsi que de géolocalisation (ici la latitude et la longitude).

En termes de paramètres, on a :

- Pour HYDROSWOT\_100m, 43277 observations (mesures réelles).
- Pour PEPSI1, 29596 observations (issues d'un modèle numérique).
- Pour PEPSI2, 37331 observations (issues d'un modèle numérique).

NB : on étudie les bases HYDRoSWOT\_100m\_corrected\_flowacc, PEPSI1\_reaches\_raw\_flowacc et PEPSI2\_reaches\_raw\_flowacc, qu'on appelle HYDRoSWOT\_100m, PEPSI1 et PEPSI2 par souci de facilité.

## 2 Vérification de la cohérence des données

On vérifiera dans cette section, la cohérence du jeu de données HYDRoSWOT\_100m en le comparant aux bases de données PEPSI1 et PEPSI2 à différentes échelles. La comparaison se fera grâce à des tests statistiques, dont on expliquera le principe au préalable. On pourra ensuite tirer de la base de données une autre base constituée uniquement des données validées par les tests.

### 2.1 Principe des tests statistiques

Lors de la comparaison de deux échantillons, on teste en premier l'égalité des variances (F-test ou test de Fisher). On teste l'égalité des moyennes dans un second temps (T-test, test de Student ou test de Welch) pour enfin conclure ou non de l'égalité des Lois de Probabilité régissant les deux échantillons. Chaque test est effectué sous R.

**Test de Fisher sous R :** Soit  $\sigma_1^2$  la variance de l'échantillon 1 de taille  $n$ , inconnue, et  $\sigma_2^2$  la variance de l'échantillon 2 de taille  $p$ , inconnue aussi. Les échantillons sont constitués d'observations d'une même variable. On veut comparer leur variances. On les estime donc à l'aide de la variance empirique,  $S_1^2$  pour l'échantillon 1, et  $S_2^2$  pour l'échantillon 2. La variance empirique suit ici une loi de khi-deux respectivement à  $n - 1$  degré de liberté et  $p - 1$  degré de liberté (Cochran). Le test de Fisher de niveau  $\alpha$  consiste à rejeter l'hypothèse nulle " $H_0 : \sigma_1^2 = \sigma_2^2$ " si la statistique de test  $F = \frac{S_1^2}{S_2^2} \frac{p}{n}$  est très petit ou très grand devant 1.  $F$  suit une loi de Fisher à  $(n - 1, p - 1)$  degrés de liberté. La règle de décision revient donc à rejeter l'hypothèse nulle si  $F$  est plus petit que le quantile de niveau  $(\alpha/2, n - 1, p - 1)$  d'une fisher à  $(n - 1, p - 1)$  degrés de liberté, ou plus grand que le quantile de niveau  $(1 - \alpha/2, n - 1, p - 1)$ . Les sorties du F-test sous R sont la valeur de F, des degrés de libertés ainsi que de la p-value (dont on ne se servira pas ici).

**Test de Student sous R :** On estime les moyennes  $\mu_1$  et  $\mu_2$  des échantillons 1 et 2 de taille  $n$  et  $p$  par leur moyennes empiriques. On souhaite savoir si elles sont égales. On teste la nouvelle hypothèse nulle  $H'_0 : \mu_1 = \mu_2$ . Nous utilisons pour cela la statistique de test  $T$ , qui est égale à la différence de ces moyennes empiriques, divisée par  $\frac{\sigma}{\sqrt{n+p-2}}$ , où la variance  $\sigma$  est égale à  $\sigma_1 = \sigma_2$  si  $H_0$  est vraie. On a donc que  $T$  suit une student à  $(n + p - 2)$  degrés de liberté. Si les variances ne sont pas égales, on pose  $d = (n + p)/2$  et on estime les variances par leur variance empiriques qui suivent cette fois une khi-deux à  $d$  degrés de liberté. On a alors que  $T$  suit une student à  $d - 1$  degré de liberté. Le T-test sous R donne la valeur de  $T$ , le degré de liberté ainsi que la p-value (dont on ne se servira pas ici). On rejette l'hypothèse d'égalité des moyennes lorsque  $T$  est loin de zéro.

## 2.2 Tests d'égalité des variances et des moyennes sur des échantillons à grande échelle géographique

Dans le but de se donner une référence, on étudie ici des comparaisons à des échelles géographiques trop larges. On va chercher à obtenir des échantillons validés par les tests, mais on va également s'assurer de la comparabilité des échantillons en réduisant l'échelle des données sur lesquelles on effectue les tests. D'autre part, les variables dont on veut le plus s'assurer de l'égalité sont les variables spatiales telles que « dH » et « W » ainsi que le débit « Q ». Les tests sont effectués sur ces trois variables seulement.

### 2.2.1 Tests de comparaison sur l'ensemble de HYDRoSWOT\_100m avec l'ensemble de PEPSI1 et PEPSI2.

Dans les Tableau 1 et Tableau 2 se trouvent les valeurs des statistiques de tests de Fisher et de Student pour les premiers tests de comparaison d'échantillon. Le Tableau 1 concerne l'ensemble des données de HYDRoSWOT\_100m (donc toutes les stations de toutes les rivières de la base de données) et PEPSI1 (donc toutes les rivières, même les non américaines). Le Tableau 2 concerne l'ensemble des données de HYDRoSWOT\_100m et PEPSI2.

On sait que l'ensemble des stations d'une multitude de rivières d'Amérique du Nord et les stations ponctuelles PEPSI ne sont pas comparables. Les tests vont le confirmer, et on obtiendra des valeurs aberrantes pour les statistiques de tests. Ces tableaux ne sont bien sûr qu'un exemple de tests non concluants.

- Les paramètres du test entre HYDRoSWOT et PEPSI1 sont :  $n = 43277$  observations HYDRoSWOT,  $p1 = 29596$  observations , l'intervalle de niveau  $\alpha = 0.05$  : [0.9793; 1.0211], dans lequel doit se situer la statistique de Fisher pour accepter l'hypothèse d'égalité des variances (dont les bornes sont les  $\alpha/2$ -quantile et  $1 - \alpha/2$ -quantile).
- Les paramètres du test entre HYDRoSWOT et PEPSI2 sont :  $n = 43277$  observations HYDRoSWOT,  $p2 = 373331$  observations PEPSI2, l'intervalle de niveau  $\alpha = 0.05$  et [0.9806; 1.0197].
- La borne de Student est environ égale à 1.96 (borne à laquelle la statistique de Student doit y être inférieure en valeur absolue pour accepter l'hypothèse d'égalité des moyennes).

Tested variable	Fisher	Student	Mean (HYDRoSWOT_100m)	Mean (PEPSI1)
dH	27.909	0.49551	1.543	1.522
W	0.0069	-61.121	132.921	779.5246
Q	0.1601	-58.245	572.1988	2447.7209

TABLE 1 – Comparison between HYDRoSWOT\_100m and PEPSI1 statistics

Tested variable	Fisher	Student	(Mean HYDRoSWOT)	Mean (PEPSI2)
dH	21.829	-2.425	1.543	1.647
W	0.0034	-63.595	132.921	984.331
Q	0.0475	-64.708	572.1988	3869.8596

TABLE 2 – Comparison between HYDRoSWOT\_100m and PEPSI2 statistics

On n'a clairement égalité des variances et moyennes pour toutes les variables, donc on rejette les hypothèses nulles  $H_0$  et  $H'_0$ . Les valeurs des statistiques de Fisher sont très éloignées de 1 pour chacune des variables. Pour « dH », la variance de la base HYDRoSWOT\_100m est plus élevée que celle de PEPSI1 et de PEPSI2; pour « W » et « Q », inversement. Les valeurs des statistiques de Student pour « W » et « Q » sont très grandes également, on a que la moyenne de chez HYDRoSWOT\_100m est bien plus petite que chez PEPSI1 ou PEPSI2. Les seules valeurs correctes pour Student sont celles des tests

sur « dH ». Cela tombe bien car les moyennes de dH sont très petites de manière générale dans cette base de données, cependant cela n'est pas du tout assez significatif pour prétendre qu'il existe une cohérence spatiale entre les bases de données.

### 2.2.2 Tests de comparaison sur un échantillon de HYDROSWOT\_100m appartenant à une rivière, avec un échantillon de PEPSI1 et PEPSI2 appartenant à la même rivière.

On réduit l'échelle une première fois : on garde chez PEPSI1 et PEPSI2 les données des stations de rivières américaines ; on garde chez HYDROSWOT\_100m les données des rivières qui existent dans la base de données des rivières américaines PEPSI.

En Figure 3 on a les résultats des tests statistiques entre ces échantillons réduits. En Figure 4, les moyennes des variables par échantillon.

River	Tested variable						Test parameters			
	dH (m)		W (m)		Q (m³/s)		Fisher	HYDROSWOT	PEPSI	Student bound = 1,97
	Fisher	Student	Fisher	Student	Fisher	Student	min	max	#observations	#observations
Ohio P1	5,69	7,689	13,76	-1,127	2,815	1,453	0.831	1,191	305	1100
Mississippi Upstream P1	4,93	9,505	16,11	-14,67	17,39	6,87	0.866	1,160	1427	486
Mississippi Downstream P1	2,327	-9,27	0,393	-52,3	2,201	-30,53	0.891	1,123	1427	972
Connecticut P1	2,034	-3,2678	0,561	-36,98	1,089	-7,204	0.8124	1,2204	266	627
Cumberland P1	0,377	-14,16	4,627	-18,41	0,757	-9,461	0.7605	1,2852	137	648
SacramentoDownstream P1	2,172	3,581	0,705	5,574	3,253	4,89	0.859	1,157	468	1386
IowaRiver P2	1,419	3,44	0,248	-4,94	8,238	6,52	0.868	1,144	471	2928
MissouriMidsection P2	2,978	14,69	1,954	-8,22	6,438	11,81	0.928	1,078	3228	2380
MissouriDownstream P2	3,898	5,39	0,977	-9,919	9,641	16,28	0.932	1,073	3228	2975
MissouriUpstream P2	2,509	14,32	0,501	-15,765	5,792	11,04	0.928	1,078	3228	2380

FIGURE 3 – Comparison between HYDROSWOT\_100m data from all stations on a river with PEPSI data from the station on the same river

River	Tested variable					
	dH (m)		W (m)		Q (m³/s)	
	mean H	mean P	mean H	mean P	mean H	mean P
Ohio P1	2,768	1,18	613,3	642,67	4554,6	4084,3
Mississippi Upstream P1	3,13	2,09	542,5	657,9	7010,2	5447,2
Mississippi P1	3,13	4,15	542,5	1338,9	7010,2	15307
Connecticut P1	0,829	1,02	168,9	569,27	355,83	634,93
SacramentoDownstream P1	1,689	1,376	118,06	106,72	348,3	272,56
IowaRiver P2	1,62	1,38	136,2	178,2	316,3	157,9
MissouriMidsection P2	2,33	1,814	225,5	242,28	1310,7	1016,7
MissouriDownstream P2	2,33	2,15	225,5	248,09	1310,7	922,03
MissouriUpstream P2	2,33	1,813	225,5	273,1	1310,7	1033,1

FIGURE 4 – Means of each variable for each sample

On compare à chaque fois un échantillon de HYDROSWOT\_100m constitué uniquement des observations d'une même rivière, par exemple Ohio, aux observations de cette même rivière dans la base PEPSI correspondante, ici PEPSI1, comme précisé dans la case du nom de la rivière. Dans le tableau intitulé « Test parameters » sont les bornes de l'intervalle dans lequel la statistique de Fisher doit se trouver pour admettre que les variances sont égales statistiquement (autrement dit,  $[\min; \max] = [\alpha/2\text{-quantile}; 1 - \alpha/2\text{-quantile}]$ ). La borne de Student, le  $\alpha/2$ -quantile de la Student, varie très peu en fonction du degré de liberté. Pour  $\alpha = 0.05$ , elle se situe toujours aux environ de 1.97. En vert sont surlignées les valeurs des statistiques de Fisher et de Student comprises dans l'intervalle, c'est-à-dire celles qui permettent d'accepter l'égalité des variances.

On constate que les valeurs prises par la statistique de test de Fisher sont bien moins aberrantes que celles des premiers tests partie 2.2.1, Tableau 1 et Tableau 2. De même, la statistique de test de Student prend des valeurs moins élevées en valeur absolue que celles des premiers test. Cependant, elles restent bien trop éloignées encore pour pouvoir affirmer égalité des moyennes. On va donc réduire l'échelle une seconde fois.

### 2.3 Tests d'égalité des variances et des moyennes sur des tronçons de 100km

On restreint la zone géographique pour laquelle on garde les échantillons HYDROSWOT\_100m. On construit alors une nouvelle base correspondant à un ensemble de stations appartenant à des tronçons de 100km de rivière comprenant une station PEPSI. Pour une rivière donnée, on choisit les stations

**HYDRoSWOT\_100m** le plus proche possible de la station PEPSI de cette rivière, le plus loin accepté étant **100km**. En priorité le choix se porte vers les stations côté embouchure de la rivière (ou le point de confluence si la rivière est un affluent). La sélection de ces stations HYDRoSWOT\_100m a été effectuée à l'aide de QGIS. Elle n'a pas été possible pour chaque rivière nord américaine de PEPSI. Du reste, on effectue les tests statistiques sur le même principe que dans la partie précédente.

### 2.3.1 Résultats des tests

Voici en Figure 5 un tableau des résultats des tests statistiques de Fisher et de Student pour les observations de HYDRoSWOT\_100m par rivière à la nouvelle échelle; et en Figure 6 le tableau des moyennes par échantillon et par variable.

River	Tested variable						Tests parameters			
	dH (m)		W (m)		Q (m³/s)		Fisher	HYDRoSWOT	PEPSI	
	Fisher	Student	Fisher	Student	Fisher	Student	min	max	#observations	#observations
Connecticut	0,922	-3,69	0,244	-27,23	1,94	-0,827	0,729	1,326	102	627
Sacramento Dstream	2,367	3,493	0,0267	-23,053	1,1512	0,6417	0,718	1,334	87	1386
SanJoaquin	3,12	2,7	0,015	-8,24	4,076	-0,736	0,614	1,646	72	60
Ohio	5,28	1,8395	1,008	0,1849	1,904	2,049	0,346	2,01	12	1100
Kanawha	4,955	3,709	38,09	-2,89	2,579	-0,207	0,6125	1,4955	43	648
IowaRiver	1,066	-1,2962	0,0296	-14,29	1,167	-0,4163	0,8233	1,1964	240	2928
Wabash	1,685	-0,2259	0,0157	-15,202	2,764	1,9614	0,785	1,254	181	648
Platte	0,77	-1,99	0,039	-23,3	0,518	1,456	0,426	1,849	17	288
Missouri Midsection	0,086	-29,394	0,0664	-6,628	0,0303	-18,142	0,4984	1,6765	23	2975
Mississippi Upstream	4,793	14,826	2,17	-11,773	6,81	6,016	0,823	1,211	361	486

FIGURE 5 – Comparison between HYDRoSWOT\_100m data between PEPSI data by river (100km scale)

River	Variables					
	dH (m)		W (m)		Q (m³/s)	
	mean H	mean P	mean H	mean P	mean H	mean P
Connecticut	0,782	1,021	263,29	569,2	573,9	634,9
Sacramento Dstream	2,073	1,376	74,86	106,72	285,14	272,55
SanJoaquin	1,871	1,306	93,8	314,4	188,9	208,2
Ohio	2,98	1,18	649,19	642,67	6004,47	4084,268
Kanawha	0,879	0,335	207,13	240,8	607,83	629,29
IowaRiver	1,279	1,383	89,27	178,22	152,85	157,96
Wabash	2,494	2,527	277,01	1405,83	1141,5	979,02
Platte	0,41	0,58	88,14	556,1	176,56	139,4
Missouri Midsection	0,5814	2,1524	214,15	248,09	615,42	922,03
Mississippi Upstream	4,795	2,0904	585,91	657,93	7106,851	5447,2

FIGURE 6 – Mean of each variable for each sample

En rouge sont surlignées les valeurs des statistiques de tests trop aberrante dans le tableau de gauche Figure 5 ; dans le tableau de droite, les nombres de données trop petits. Sont surlignées en vert les valeurs qui statistiquement permettent d'accepter avec conviction que les moyennes sont égales pour Student, ou les variances pour Fisher.

- On voit que la plupart des tests donnent égalités des moyennes pour le débit Q. Pour Ohio, la statistique de Student est à quelques centièmes d'être comprise dans l'intervalle mais on remarque que les variances sont égales, on peut donc considérer le test comme satisfaisant et ce même pour le petit nombre de données disponible pour Ohio.
- Missouri Midsection et Mississippi Upstream sont les seules rivières dont aucun des tests ne confirme l'égalité des variances ou des moyennes. Pour Missouri, il y a peu de données et en plus de cela elles ne sont pas représentatives (sinon on aurait au moins égalité des variances). Le **Mississippi est une très grande rivière et malheureusement la seule station HYDRoSWOT\_100m aux alentours de la station PEPSI Mississippi Upstream était assez éloignée** (Remarque : en général il y a un cruel manque de données sur l'ensemble du Mississippi).
- On constate en effectuant les tests qu'on perd en précision sur la largeur W lorsqu'on en augmente sur le débit Q.
- Les valeurs des statistiques de tests pour dH sont assez satisfaisantes. On sait également que l'erreur de mesure sur W est assez grande (en mètres). Au vu des résultats pour dH et Q et du tableau en Figure 6, à ce niveau là, on peut se dire sans trop de risque que les données de HYDRoSWOT\_100m tronc sont cohérentes.

### 2.3.2 Définition d'une base de données constituée des échantillons validés par les tests

Voici en Figure 7, 8 et 9 la représentation en chevelu par QGIS des rivières Connecticut, Iowa, Kanawha, Ohio, Sacramento Downstream, San Joaquin et Wabash. Ce sont, comme vu précédemment,

les rivières dont les tests sur les échantillons de tronçons de 100km sont positifs et sur lesquelles se trouvent les stations HYDRoSWOT\_100m associés.

Toutes ces stations de tronçons de 100km de rivières (à l'exception de celles appartenant au Missouri, Mississippi et Platte) sont extraites dans une base de données nommée HYDRoSWOT\_100m\_tronc. Cette base est considérée valide. Les stations en jaune sont les stations HYDRoSWOT\_100m\_tronc de la rivière, en rouge ou en vert, la station PEPSI de la rivière. En bleu, ce sont les stations HYDRoSWOT\_100m n'appartenant pas aux tronçons.

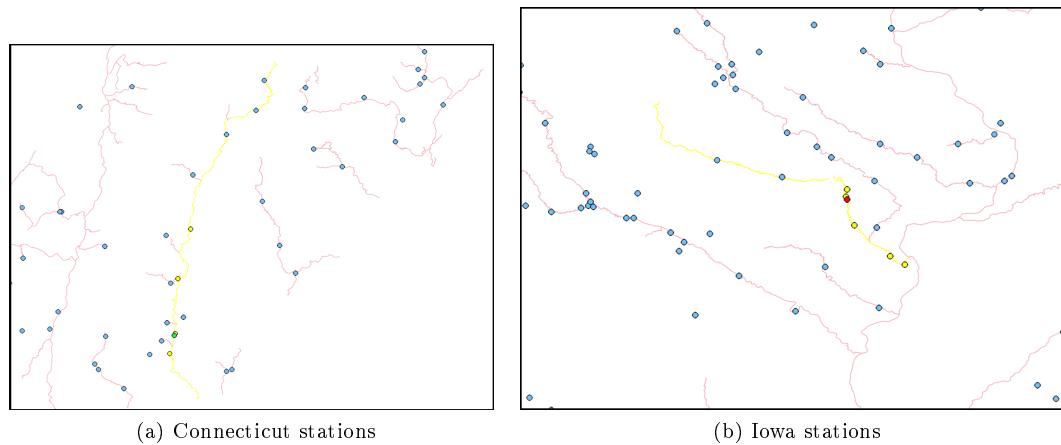


FIGURE 7 – HYDRoSWOT\_100m\_tronc and PEPSI stations (QGIS) - Connecticut & Iowa

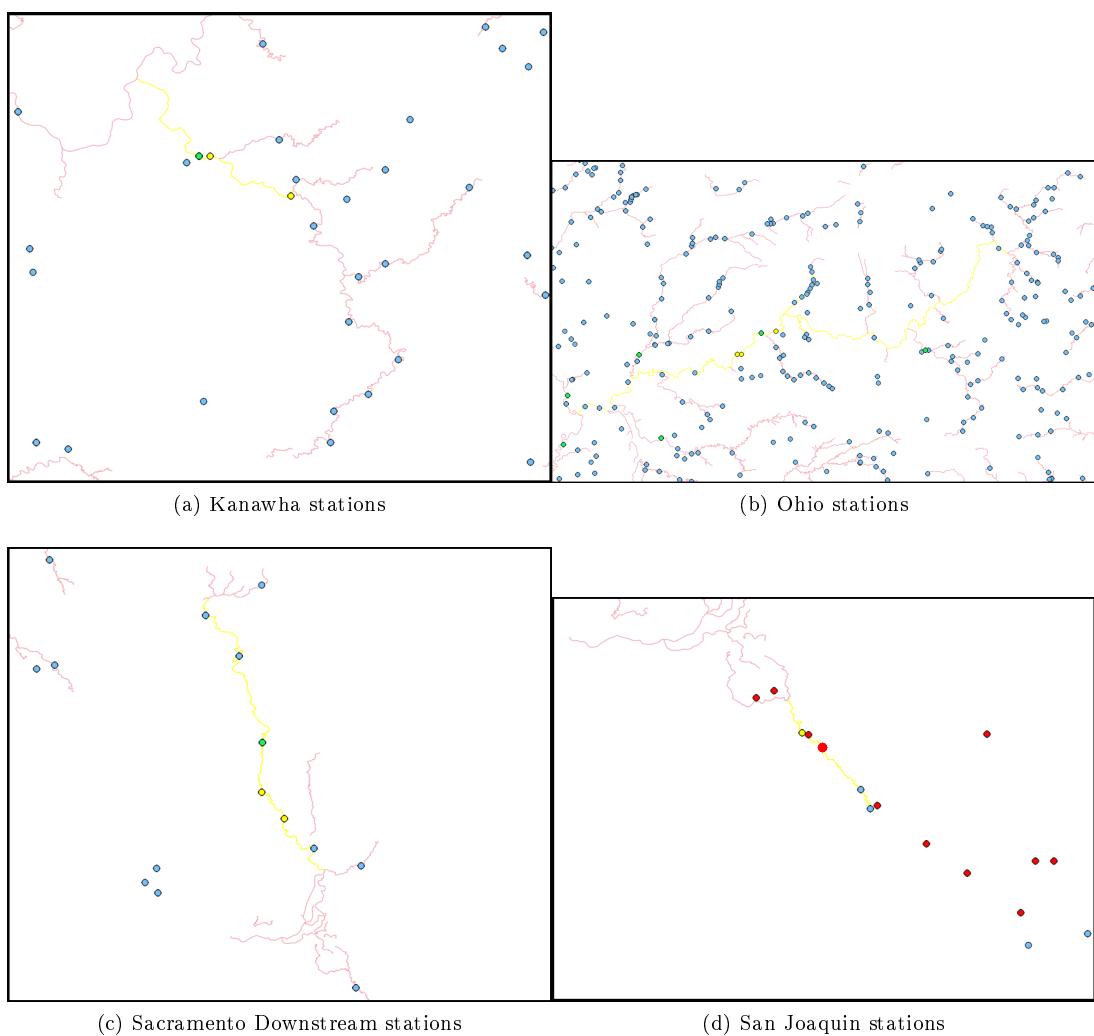


FIGURE 8 – HYDRoSWOT\_100m\_tronc and PEPSI stations (QGIS) - Kanawha & Ohio ; Sacramento Downstream & San Joaquin

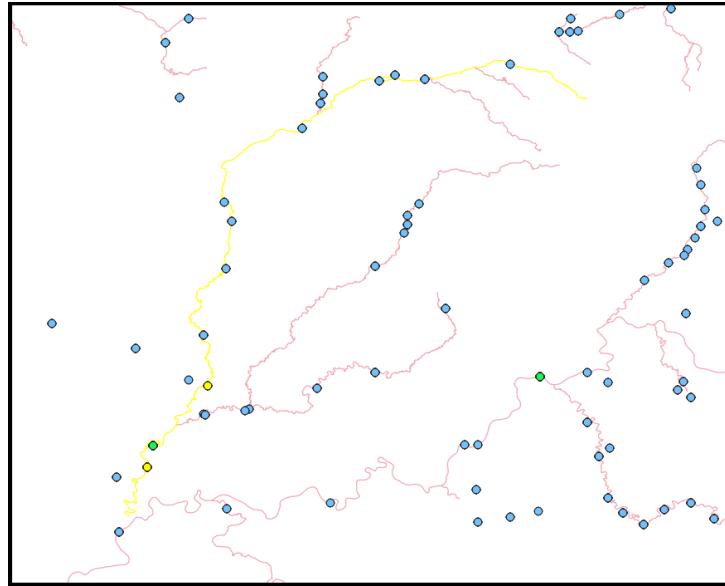


FIGURE 9 – HYDROSWOT\_100m\_tronc and PEPSI stations (QGIS) - Wabash

Dans la suite, les analyses sont uniquement faites sur HYDROSWOT\_100m\_tronc. Les conclusions tirées des analyses sont valides dans ce cadre et non sur l'ensemble des données ou sur des données à une échelle différentes. Par souci de cohérence et simplicité, on crée également une base de données PEPSI restreinte à PEPSI\_tronc qui correspond aux données PEPSI des 7 rivières sélectionnées.

### 3 Analyse statistique descriptive de la base de données constituées des mesures de tronçons de 100km

#### 3.1 Analyse de la base de données HYDROSWOT\_100m\_tronc et vérification avec PEPSI\_tronc

On va effectuer plusieurs analyses de la base de données HYDROSWOT\_100m\_tronc en parallèle à PEPSI\_tronc, afin d'obtenir un résumé synthétique de la base de données, une classification des rivières nord américaine. Les méthodes employées sont :

1. L'Analyse des corrélations entre les variables
2. L'Analyse en Composantes Principales (ACP)
3. L'Analyse par Classification Ascendante Hiérarchique (CAH ou clustering)

##### 3.1.1 Analyse de la matrice des corrélations des variables, comparaison avec la matrice de corrélation de PEPSI\_tronc

On analyse ici les corrélations entre les variables de HYDROSWOT\_100m\_tronc puis de PEPSI\_tronc. On va ainsi pouvoir voir apparaître des liens entre plusieurs variables qui vont transparaître dans l'ACP et l'ACH.

**a. Corrélation HYDROSWOT\_100m\_tronc** Voici en Figure 10 la matrice des corrélations entre les variables. Plus la couleur de l'ellipse s'intensifie et plus elle est fine, plus la corrélation est forte. Si la couleur est bleue, la corrélation est positive; si elle est rouge alors elle est négative. Si elle est pâle ou blanche, alors elle est nulle ou presque. En Figure 11 se trouve la même matrice avec les valeurs numériques.

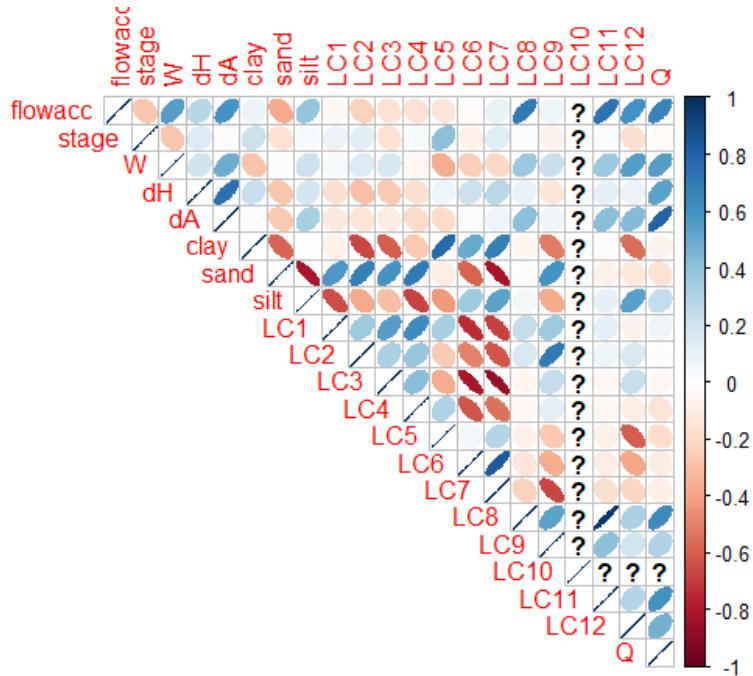


FIGURE 10 – Correlation matrix for HYDROSWOT\_100m\_tronc variables

- Corrélations positives
  - dA et Q, dA et dH, Q et dH : forte corrélation entre le débit et les variations d'élévations et d'aire. On a une forte dépendance linéaire entre ces trois variables.
  - flowacc et Q, dH et W : l'aire draînée est fortement corrélée au débit, à la variation d'élévation et la largeur de la rivière.
  - W et Q, flowacc : si l'aire draînée augmente, la largeur augmente, de même pour le débit
  - LC11, LC12 et Q ainsi que dA sont très fortement corrélés entre eux.
- On remarque ainsi un groupe de variable fortement corrélées entre elles : Q, flowacc, dA, dH, W, ainsi que LC11 et LC12 (le taux de terre stériles et le taux de présence d'eau). Ces variables évoluent toutes dans le même sens. Les autres corrélations positives de ces variables là sont trop faibles pour être significatives ( $< 0.35$ ).

  - sand est fortement corrélée avec chacune des variables LC1, LC2, LC3, LC4, LC9. Le taux de sable augmente en même temps que le taux d'arbres (LC1 à LC4) augmente. LC9 correspond au taux de présence de construction humaine.
  - de même clay et LC5, LC6, LC7 sont corrélés. Le taux d'argile dans la terre est linéairement dépendant du taux d'herbacées (LC6) d'arbustes (LC5) et de terres cultivées (LC7).

- Corrélations négatives
  - silt et sand, LC1, LC2, LC3, LC4 : le taux de limon est fortement corrélé négativement avec le groupe de variables du sable et des arbres.
  - clay et sand, LC1, LC2, LC3, LC4 : le taux d'argile augmente quand le taux de sable et d'arbre diminue
  - LC6, LC7 et LC1, LC2, LC3, LC4 : le taux d'herbacées et de terres cultivées sont fortement corrélés négativement avec le taux d'arbres. Plus il y a d'arbres, moins il y a d'herbacées et de terres cultivées.
- On remarque que le groupe de variable de sable et d'arbres, (sand LC1, L2, LC3 et LC4) s'oppose au groupe de variable d'argile, boue, herbacées et terres cultivées (silt, clay, LC6 et LC7). On remarque également qu'il n'y a pas de corrélations notable avec le groupe de variable Q, flowacc, W, dA, dH, LC11 et LC12. Cela signifie que le débit, l'aire draînée etc évoluent indépendamment des autres groupes de variables. Les seules variables qui influent sur le débit sont celles corrélées positivement avec ce dernier.

	flowacc	stage	W	dH	dA	clay	sand	silt	LC1	LC2	LC3	LC4	LC5	LC6	LC7	LC8	LC9	LC10	LC11	LC12	Q
flowacc	1,0000	-0,2748	0,5538	0,2812	0,5991	0,0886	-0,3726	0,3975	-0,0333	-0,2299	-0,1556	-0,1591	-0,1312	-0,0295	0,0970	0,7084	0,0780	NA	0,7374	0,6061	0,6684
stage	1,0000	-0,2744	0,1451	-0,0180	0,2180	-0,1563	0,0375	0,0829	0,1297	-0,1558	0,0352	0,4123	-0,0770	0,1360	0,0230	-0,0613	NA	0,0216	-0,1608	-0,0220	
W	1,0000	0,2030	0,4941	-0,2847	-0,0120	0,2176	0,0490	0,1494	0,1713	-0,0478	-0,3641	-0,2472	-0,2034	0,3742	0,2212	NA	0,3718	0,5591	0,5520		
dH	1,0000	0,7413	0,2223	-0,2775	0,1845	-0,1633	-0,2932	-0,2615	-0,1649	0,0738	0,2162	0,2704	0,0948	-0,1378	NA	0,1050	0,0822	0,5302			
dA	1,0000	-0,0157	-0,2626	0,3556	-0,1139	-0,1405	-0,1066	-0,1969	-0,1930	0,0269	0,0798	0,4156	0,0686	NA	0,4275	0,4377	0,8083				
clay	1,0000	-0,5823	0,0092	-0,0750	-0,6607	-0,6015	-0,2695	0,7722	0,5099	0,6818	-0,0552	-0,5110	NA	-0,0191	-0,5590	-0,0696					
sand	1,0000	0,8183	0,5724	0,6845	0,5931	0,7055	-0,0969	-0,5878	-0,8249	-0,0144	0,5946	NA	-0,0738	-0,1235	-0,1551						
silt	1,0000	-0,6496	-0,3758	-0,3019	-0,6768	-0,4269	0,3599	0,5313	0,0581	0,3706	NA	0,1055	0,5492	0,2415							
LC1	1,0000	0,3683	0,5513	0,6245	0,3329	-0,7457	-0,6889	0,2408	0,3668	NA	0,1190	-0,0699	0,0788								
LC2	1,0000	0,3274	0,3849	-0,2572	-0,4912	-0,6226	0,1672	0,0790	NA	0,0749	0,1616	0,0113									
LC3	1,0000	0,4213	-0,3608	-0,8194	-0,8627	-0,0544	0,2413	NA	-0,0425	0,2397	-0,0464										
LC4	1,0000	0,3128	-0,6274	-0,5410	-0,0336	0,1176	NA	-0,0479	-0,0942	-0,1470											
LCS	1,0000	0,0598	0,2915	-0,0788	-0,2646	NA	-0,0851	-0,6046	-0,1813												
LC6	1,0000	0,8369	-0,1384	-0,3667	NA	-0,0904	-0,3812	-0,1082													
LC7	1,0000	-0,2223	-0,6615	NA	-0,1660	-0,2115	-0,0862														
LC8	1,0000	0,5370	NA	0,9659	0,3240	0,6386															
LC9	1,0000	NA	0,4161	0,1901	0,3050																
LC10	1,0000	NA	NA	NA	NA																
LC11	1,0000	NA	NA	NA	NA																
LC12	1,0000	NA	NA	NA	NA																
Q	1,0000																				

FIGURE 11 – Numerical correlation matrix for HYDROSWOT\_100m\_tronc variables

**b. Corrélation PEPSI\_tronc** On s'intéresse maintenant à la matrice des corrélations de PEPSI\_tronc, Figure 12. Les valeurs numériques sont dans la matrice Figure 13.

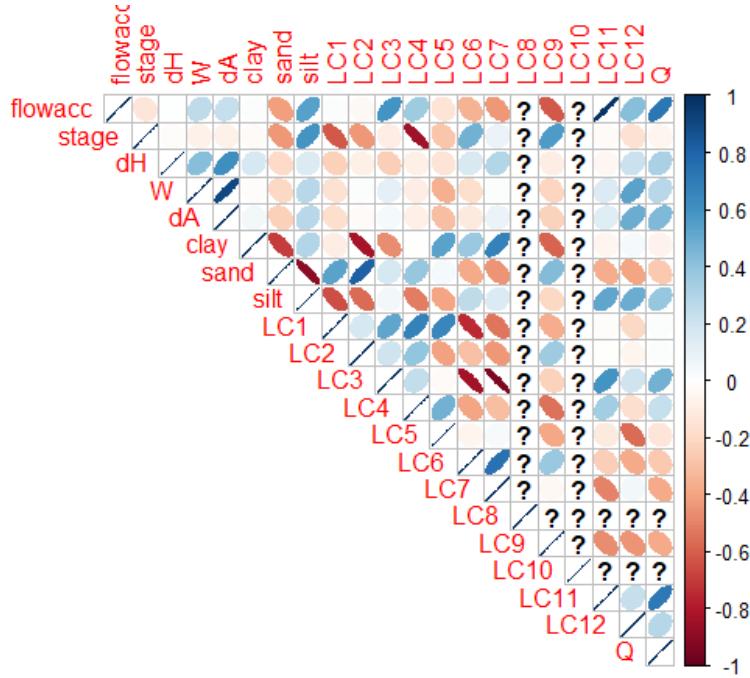


FIGURE 12 – Correlation matrix for PEPSI\_tronc

— Corrélations positives : Par souci de simplicité, on va noter que A est corrélée avec B,C ; qui signifie que A est corrélée avec B et A est corrélée avec C.

— flowacc est corrélée avec silt, LC11, LC12, Q

— dH avec W, dA

— W avec dA, LC12

— dA avec LC12, Q

— silt avecs LC11, LC12, Q

Le même groupe de variables corrélées positivement entre elles est présent : Q, flowacc, dH, dA, W, LC11 et LC12. A ce groupe s'ajoute silt qui est moyennement corrélé à toutes ces variables.

— stage et silt, LC6

— clay et LC5,LC6,LC7

— sand et LC1, LC2, LC3

Le groupe de sable et d'arbre se retrouve encore ici, sand, LC1, LC2, LC3 et LC4 avec une corrélation moins forte pour LC3 et LC4. Le groupe clay, silt, LC5, LC6, LC7 est complété ici par stage. L'élévation est dans la base de données PEPSI\_tronc, corrélée positivement avec le taux de limon et d'herbacées.

— Corrélations négatives :

— clay et sand, LC2, LC3, LC9

— LC1, LC2, LC3, LC4 et LC5, LC6.

Encore une fois on a le groupe de variables des arbres et du sable, sand LC1,LC2, LC3, et LC4 qui s'oppose au groupe d'argile et d'herbacées, clay, LC5, LC6 et LC7. On a également encore une fois des corrélations nulles ou quasi-nulles entre Q et LC1,LC2,LC3,LC4, LC5, LC6. Cela signifie que pour PEPSI aussi, les variables Q, flowacc, etc n'ont d'influences que les unes sur les autres, et évoluent indépendamment des groupes de variables d'argile, de sable et de végétations.

	flowacc	stage	dH	W	dA	clay	sand	silt	LC1	LC2	LC3	LC4	LC5	LC6	LC7	LC8	LC9	LC10	LC11	LC12	Q
flowacc	1,0000	-0,1396	0,0010	0,2560	0,2305	0,0180	-0,4183	0,5461	0,0160	-0,0371	0,5973	0,3531	-0,1439	-0,3424	-0,4358	NA	-0,6131	NA	0,9715	0,4286	0,7206
stage		1,0000	-0,0187	-0,0724	-0,0709	-0,0267	-0,4216	0,5942	-0,6162	-0,4395	-0,0928	-0,8465	-0,2761	0,4780	0,0880	NA	0,5623	NA	-0,0269	-0,1551	-0,0540
dH			1,0000	0,4297	0,6168	0,1733	-0,1980	0,1520	-0,2325	-0,0815	-0,2436	-0,0835	-0,1452	0,1686	0,2996	NA	-0,0980	NA	-0,0489	0,2158	0,3269
W				1,0000	0,9013	-0,0167	-0,2024	0,2764	-0,1571	0,0154	0,1108	-0,0950	-0,3571	-0,1751	0,0148	NA	-0,2103	NA	0,1591	0,5326	0,2709
dA					1,0000	0,0507	-0,2335	0,2753	-0,1733	-0,0279	0,0402	-0,0837	-0,3088	-0,1106	0,0847	NA	-0,2267	NA	0,1364	0,4955	0,4497
clay						1,0000	-0,6955	0,2931	-0,0944	-0,8260	-0,4621	-0,0066	0,5348	0,3782	0,6746	NA	-0,5892	NA	-0,0565	0,0425	-0,0629
sand							1,0000	-0,8908	0,5334	0,8143	0,1752	0,3890	0,0469	-0,3700	-0,4403	NA	0,4320	NA	-0,3641	-0,3946	-0,2614
silt								1,0000	-0,6498	-0,5623	0,0597	-0,5123	-0,3968	0,2562	0,1572	NA	-0,2002	NA	0,5229	0,4912	0,3884
LC1									1,0000	0,1769	0,5214	0,6705	0,6516	-0,7460	-0,5350	NA	-0,3632	NA	-0,0185	-0,2087	0,0111
LC2										1,0000	0,2011	0,3931	-0,4076	-0,2958	-0,4375	NA	0,3562	NA	0,0000	-0,0586	0,0147
LC3											1,0000	0,2466	-0,0257	-0,8339	-0,9369	NA	-0,2221	NA	0,5937	0,2052	0,4780
LC4												1,0000	0,4742	-0,3941	-0,3034	NA	-0,5418	NA	0,3412	-0,1757	0,2316
LC5													1,0000	-0,0586	0,0341	NA	-0,3813	NA	-0,1135	-0,5680	-0,1367
LC6														1,0000	0,7437	NA	0,3715	NA	-0,2402	-0,3736	-0,2656
LC7															1,0000	NA	-0,0384	NA	-0,4942	0,0536	-0,3788
LC8																###	NA	NA	NA	NA	
LC9																	1,0000	NA	-0,4662	-0,4495	-0,3792
LC10																		####	NA	NA	
LC11																			1,0000	0,2310	0,7133
LC12																				1,0000	0,2840
Q																					1,0000

FIGURE 13 – Numerical correlation matrix for PEPSI\_tronc

**Remarque :** Pour PEPSI ou HYDRoSWOT, on a pas assez de données différentes de 0 de LC8 et LC10 pour dire quoique ce soit de ces variables.

### 3.1.2 Analyse en Composantes Principales de HYDRoSWOT\_100m\_tronc et de PEPSI\_tronc

**Principe de l'ACP :** Lorsque comme ici on étudie une base de données constituée d'une multitude de variables, il peut être difficile d'en avoir une vision synthétique. On a besoin d'outils pour dégager les grandes tendances des individus de la population étudiée (dans le cas présent, les rivières). Plus particulièrement, on cherche à expliquer l'inertie totale en s'intéressant à la contribution de groupe de variables. On travaille ici sur la base de données centrée réduite, avec la totalité des variables sauf « meandwave » et « sinuosity ». On compare ainsi les individus de la population par rapport à un individu de référence qui est l'individu moyen. On réduit car les variables sont d'unités différentes ; cela permet aussi d'obtenir des distances normalisées et donc de mieux visualiser les informations. De plus, en ACP centrée réduite, étudier l'inertie revient à étudier la variabilité entre les individus et l'individu moyen. Le nombre de variable étant grand, on explique l'inertie à l'aide de méta-variables, les composantes principales, qui sont des combinaisons linéaires des variables initiales. Elles sont construites en calculant les vecteurs propres de la matrice de travail, et la valeur propre associée correspond à la variance expliquée par cette combinaison linéaire de variable. On ne retient que les composantes qui expliquent le plus grand pourcentage de variabilité, pour être le plus synthétique possible.

**a. ACP HYDRoSWOT\_100m\_tronc** Voici en Figure 14 le diagramme de Pareto de la base de données centrée réduite. Ce diagramme a pour ordonnée le pourcentage de variabilité expliquée par la composante, en abscisse. En Tableau 3 le pourcentage précis de variance expliquée par composante, et la valeur propre associée à la composante.

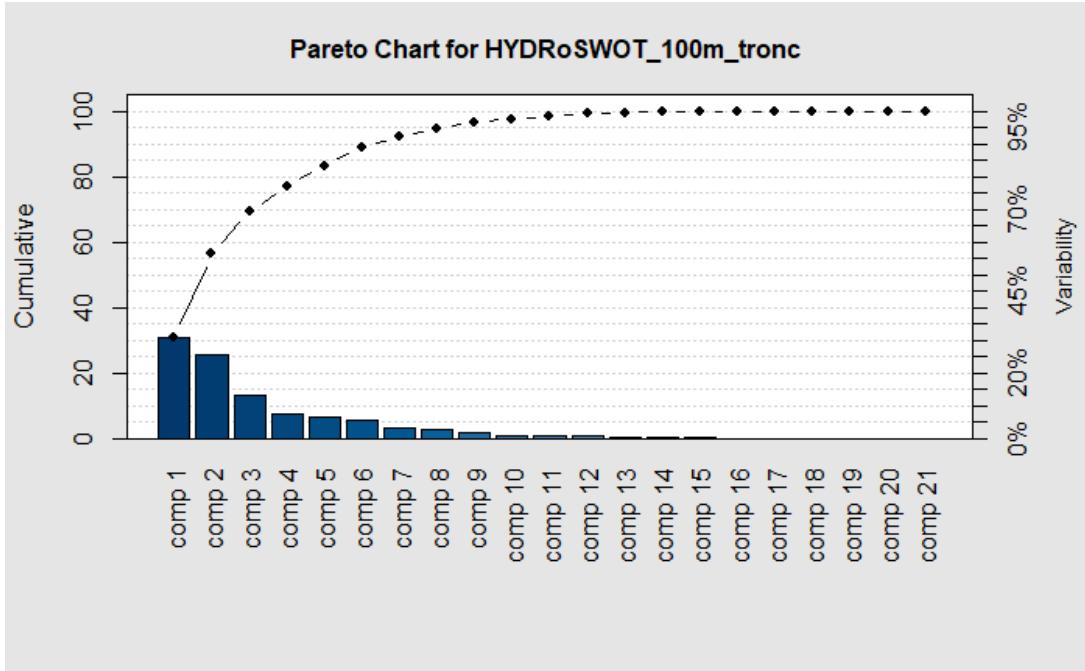


FIGURE 14 – Pareto Chart - Variability percentage per component for HYDROSWOT \_100m valid stations

	eigen value	percentage of variance
comp 1	6.167414	30.837070
comp 2	5.087610	25.438051
comp 3	2.612224	13.061122
comp 4	1.519149	7.595746
comp 5	1.260473	6.302366
comp 5	1.137726	5.688632

TABLE 3 – Percentage of variability per component

On voit que la composante 1 et la composante 2 ensemble expliquent environ 55% de la variabilité. On retient ces deux composantes et dans cette mesure, on peut constituer un nouveau plan, le plan factoriel, sur lequel on pourra projeter les individus graphiquement et faire l’interprétation.

Voici en Figure 15 le graphique obtenu sous R avec le package « FactoMineR » de projection des variables sur le premier plan factoriel, et à droite la contribution exacte des variables par dimension.

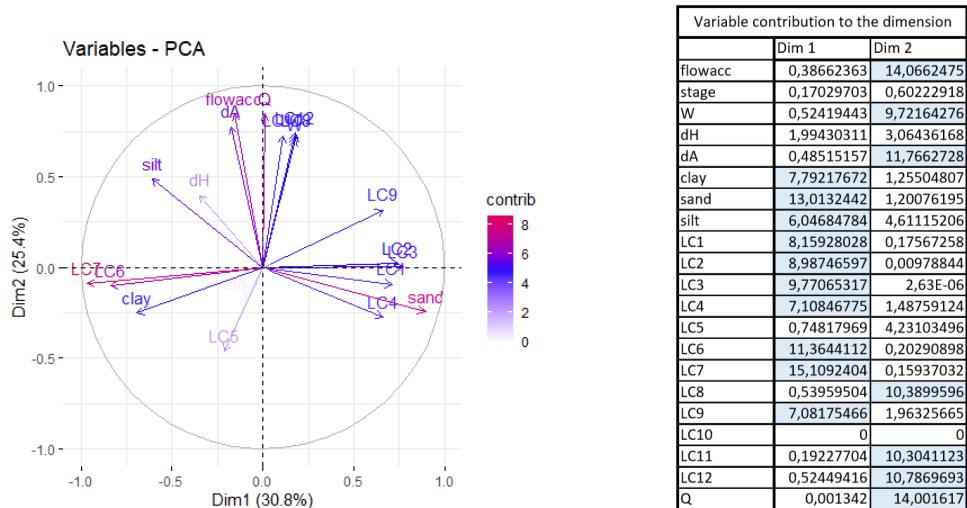


FIGURE 15 – HYDROSWOT \_100m \_tronc PCA Variable plot

**Explication du diagramme :** Le diagramme de projection des variables permet ici d'évaluer la contribution des variables aux composantes et ainsi de catégoriser les individus par groupes de variables. L'échelle de couleur donne l'importance de la contribution de la variable à l'axe. Pour plus de précision, on peut voir sur le tableau à droite les valeurs exactes de la contribution de chaque variable. Sans regarder la projection dans le plan factoriel, on considère que la contribution d'une variable est importante si elle est supérieure à la contribution moyenne, 1 sur le nombre de variables, soit  $\frac{1}{21}$  environ égal à 4,7%.

### Interprétation des composantes :

- Pour la dimension 2, les variables contribuantes sont Q, flowacc, W, LC12, LC8, LC11, dA, dH, stage, LC5 et silt. Les variables les plus importantes sont Q et flowacc, soit le débit et l'aire draînée, puis viennent dA et W. LC8 représente le taux de « terres innondables » ; LC11 le taux de « terre non cultivables », LC12 le taux de « présence d'eau » et sont moyennement importants. Typiquement, des individus qui se situent sur l'axe 1 dans les positifs se caractérisent par une valeur élevée des variables contribuantes . C'est surtout Q et flowacc les facteurs déterminant leur appartenance à cette zone du plan, plutôt que dH qui contribue très peu. Au contraire, ceux qui se trouvent dans les négatifs auront plutôt une valeur faible de Q et flowacc. En d'autres termes, les individus se trouvant sur le côté positif de la dimension 2 seront caractérisés par un haut débit et une grande aire draînée, et ceux se trouvant sur le côté négatif un débit plus faible.
- Pour la dimension 1, les variables contribuantes sont sand, LC1, LC2, LC3, LC4, du côté positif et LC7, LC6, clay du côté négatif. Les variables les plus contribuantes sont sand, LC7, LC6, et les autres sont toutes d'une importance moyenne. Les individus positionnés sur le côté positif de la dimension 2 sont caractérisés par un taux de sable élevé, ainsi que la présence de forêt (LC1, LC2, LC3 et LC4 sont différents types d'arbres) ; mais par un taux d'herbacées et de terres cultivées plus faible (resp. LC6 et LC7) ainsi qu'un taux d'argile plus faible.
- Le fait que silt et LC9 se trouvent en diagonale des deux axes signifie que ces variables sont corrélées aux variables de chaque axe, donc ne peuvent pas être incluses catégoriquement dans un seul axe. D'autre part, **la variabilité du débit n'est pas significativement influencée par les valeurs prises par les variables de la dimension 2**. Autrement dit, Q, flowacc ne sont pas corrélées au variable de la dimension 2. Que des individus se trouvent dans la partie en haut à droite du plan factoriel, ne signifie seulement qu'ils ont un fort débit et donc une forte aire draînée, et de manière indépendante un taux de sable conséquent. (**est-ce que c'est clair ou pas....**)

On remarque que les composantes correspondent exactement aux groupes de variables corrélées entre elles, vues dans la partie précédente (3.1.1).

Maintenant, on analyse en parallèle des variables, les individus (les rivières). Voici en Figure 16 la projection des individus selon leur force représentative (à gauche) et le biplot (à droite) qui correspond à la projection des individus combinée à celle des variables.

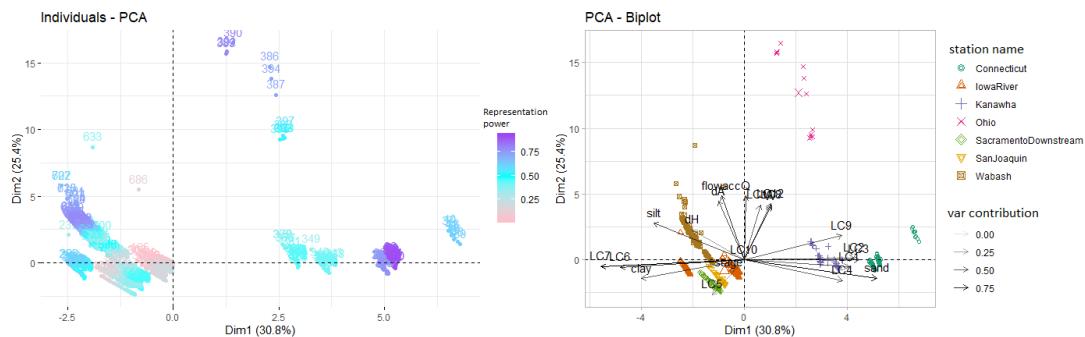


FIGURE 16 – HYDRoSWOT\_100m\_tronc PCA Individual plot by contribution (left) and Biplot (right)

**Explication des diagrammes :** Chaque observation (individu) associées à une rivière sont d'une même couleur. On peut ainsi catégoriser les rivières à l'aide des composantes puis vérifier la qualité de cette catégorisation à l'aide du diagramme de gauche. Voici en Tableau X, la force de représentation moyenne par rivière précise à gauche, ainsi que le biplot de l'ACP à droite. Plus la rivière est bien représentée par la dimension qui la caractérise, plus sa force de représentation est proche de 1. Cela correspond à l'échelle de couleur sur le diagramme de projection d'individus de gauche.

### Description des rivières :

- La rivière Ohio est située dans la partie positive de l'axe 1 et de l'axe 2. Elle est plutôt bien représentée par les composantes, selon le diagramme de gauche, et ce même si il n'y a que 12

données. Donc l'Ohio est une rivière large caractérisée par un débit élevé et une aire draînée élevée.

- La rivière Kanawha est située sur l'axe 1 côté positif. Elle est plutôt caractérisée par une forte présence de sable et d'arbres. Ceci dit, elle n'a pas un débit assez fort pour être dans les positifs de l'axe 2. La force de représentation est moyenne, donc on peut considérer que cette caractérisation est correcte.
- La rivière Connecticut est du même type que Kanawha, c'est à dire sableuse et avec une forte forêt. Cette rivière est la mieux représentée par les composantes principales.
- Wabash se situe sur les positifs de l'axe 2 et les négatifs de l'axe 1. C'est donc une rivière avec un débit élevé, une forte aire draînée, assez large mais également limoneuse, avec un taux d'argile moyen, et un taux de présence de terres cultivées et d'herbacées autour moyen
- Sacramento Downstream et San Joaquin sont assez proche. Ce sont des rivières assez argileuse, avec les herbacées et les terres cultivées. Leur force de représentation est plutôt faible.
- Iowa est pareil mais c'est la rivière la moins bien représentée.

**b. ACP PEPSI\_tronc** On prend maintenant de PEPSI les données concernant uniquement les rivières dont on a effectué l'ACP. On effectue l'ACP également dessus et on compare avec celle de HYDROSWOT. On devrait obtenir une catégorisation des rivières assez proche. Voici en Figure 17 la projection des variables sur le premier plan factoriel et en Tableau 4 le pourcentage de variance expliquée par composante avec la valeur propre associée.

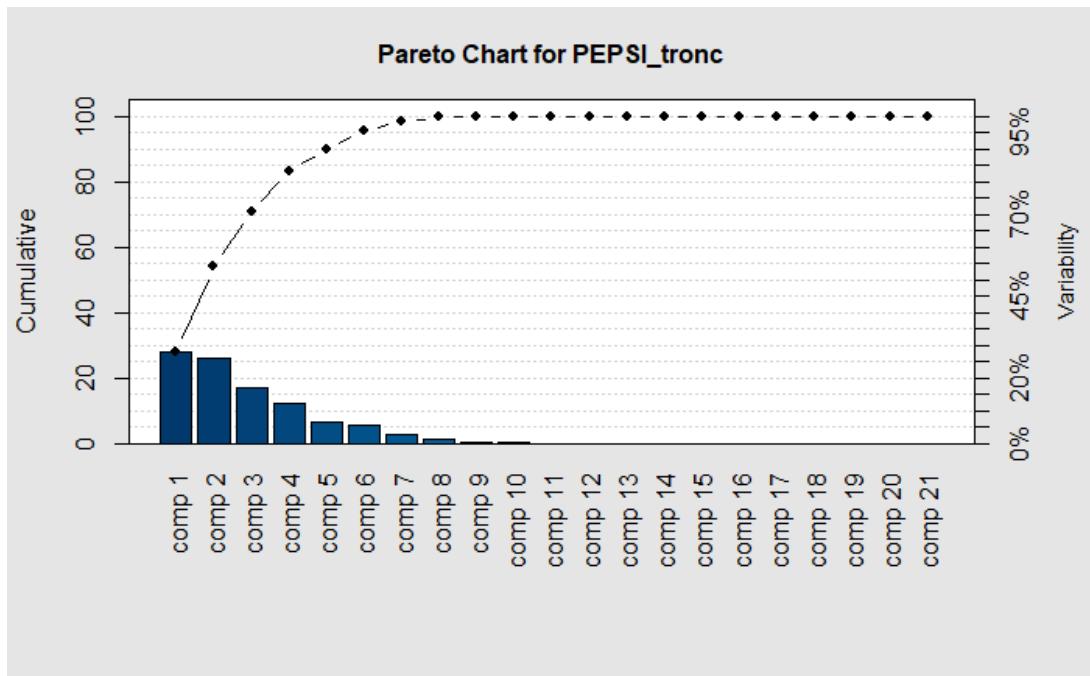


FIGURE 17 – PEPSI\_tronc Pareto Chart

	eigen value	percentage of variance
comp 1	5.296350	27.875526
comp 2	4.947272	26.038271
comp 3	3.181748	16.746045
comp 4	2.361232	12.427536
comp 5	1.257081	6.616218
comp 5	1.098338	5.780727

TABLE 4 – Percentage of variability per component

On choisit les deux premières composantes qui ici expliquent 53% de la variabilité environ. Voici en Figure 18 la projection des variables sur le premier plan factoriel de l'ACP.

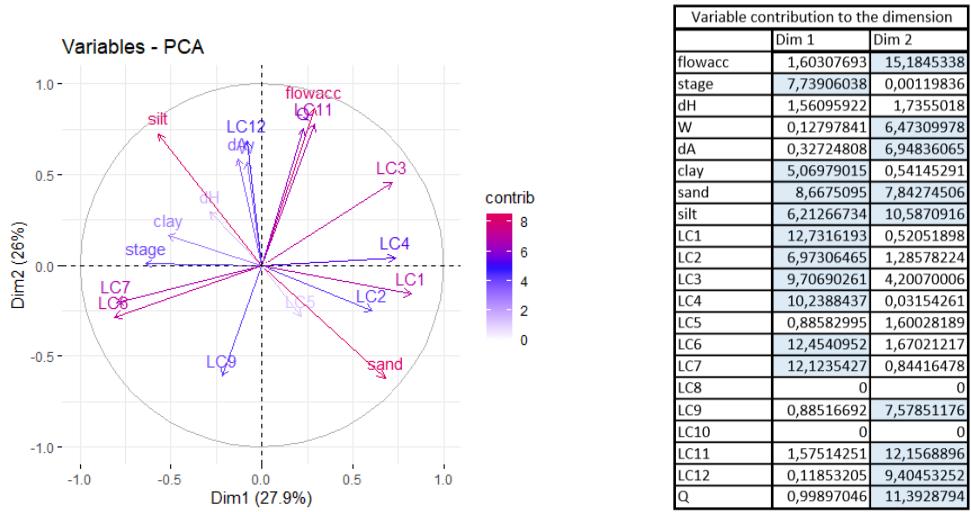


FIGURE 18 – PEPSI\_tronc PCA Variable plot

Interprétation des composantes :

- La dimension 1 constitue 28% de la variabilité entre les individus et l'individu moyen.
- Les variables qui y contribuent le plus sont, pour les positifs : LC1, LC3, LC4, LC2.
- Pour les négatifs, on a LC6, LC7, stage, clay. On remarque que stage contribue de manière notable pour la composante 1 de cette base de données comparée à l'autre ACP.
- La dimension 2 constitue 26% de la variabilité entre les individus et l'individu moyen.
- Pour les positifs, la dimension 2 regroupe : flowacc, Q, LC11, LC12, W, et dA
- LC9 pour les négatifs.
- On a silt et sand qui contribuent à la dimension 1 et à la dimension 2.

Les deux composantes ici sont assez proche de celles de l'autre ACP mais on est un peu plus nuancé, notamment avec sand, silt et LC9. On obtient les mêmes différences qu'avec l'analyses des corrélations, et les mêmes groupes de variables. Voici en Figure 19 la projection des individus sur le premier plan factoriel, ainsi que le biplot de l'ACP.

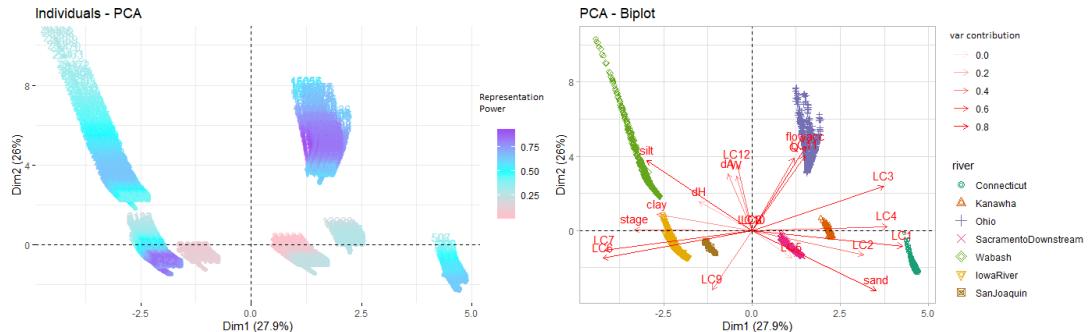


FIGURE 19 – PEPSI\_tronc PCA Individual Plot by contribution (left) and Biplot (right)

Sur le biplot, on distingue clairement les différentes rivières entre elles, pas de nuage de points entremêlés, comme on pouvait s'y attendre pour la base de données PEPSI\_tronc. On peut procéder à la description des rivières :

- La rivière Ohio est ici placée au même endroit que dans l'autre ACP : dans la zone caractérisée par un fort débit, une forte aire draînée. Ce qui confirme la représentativité des données pour le tronçon Ohio de la base de données HYDRoSWOT\_100m\_tronc.
- La rivière Connecticut est aussi à peu près au même endroit, quoique dans les négatifs de la dimension 2. Cela signifie que c'est une rivière sableuse, près de fôrets, avec un débit et une aire draînée assez faible.
- Kanawha se trouve au même endroit que sur l'autre ACP. C'est une rivière sableuse, avec bcp d'arbres, un débit moyen. Sa force de représentation est pluôt bof ici.
- Sacramento Downstream est par contre ici caractérisé par du sable et des forêts, l'opposé de ce qui était décrit dans l'autre ACP. Cette représentation est tout de même très mauvaise selon le diagramme de gauche.

- Wabash est représentée par un débit très élevé comme pour l'autre ACP, avec beaucoup de limon et d'argile
- Iowa river est argileux, ici il est aussi décrit comme avec une grande élévation et a une bonne force de représentation, avec un nb de donnée très grand
- San Joaquin est très mal représenté ici par les composantes. Il est décrit comme dans l'autre ACP, donc on en déduit que cette mauvaise force de représentation ici est due au manque de données.

**Remarque :** La description des rivières de la base de données PEPSI\_tronc faite à l'aide de l'ACP est similaire à celle que l'on obtient de HYDRoSWOT\_100m\_tronc, à l'exception de la rivière Sacramento Downstream. Malgré les deux descriptions obtenues, elle reste mal représentées par les composantes, donc par les variables du jeu de données. (Peut être un autre jeu de variable serait plus déterminant pour cette rivière)

### 3.1.3 Analyse par Classification Ascendante hiérarchique ou Clustering de HYDRoSWOT\_100m\_tronc

**Principe de l'Analyse par Classification Hiérarchique (Clustering)** De même que pour l'ACP, l'Analyse par Classification Hiérarchique ou Clustering permet de synthétiser les informations d'une base de données contenant une multitude de variables. On calcule cette fois la distance euclidienne entre tous les individus de la base de données. En fonction de quel degré de similitude on veut entre les individus, on peut les regrouper par cluster. Autrement dit, on regroupe les individus les plus similaires ie ceux dont la distance euclidienne entre eux est la plus petite, et cela permet de créer un certain nombre de groupes, les clusters. Contrairement à l'ACP où on construit les catégories et on positionne les individus, le Clustering permet de regrouper les individus pour ensuite créer des catégories. A l'aide des distances euclidiennes, on peut construire un dendrogramme qui permet de visualiser la proximité des individus et des groupes. Le dendrogramme se lit ainsi : sur l'axe des ordonnées, on lit le niveau de similarité entre deux individus. Plus le noeud de liaison entre deux individus est à une ordonnée élevée, moins ceux-ci seront similaires. Les individus se lisent en abscisses. Pour choisir le bon nombre de groupes et obtenir un bon partitionnement de la base de données, on va tenter de minimiser l'inertie à l'intérieur des groupes et de la maximiser entre les groupes, pour obtenir des groupes bien différenciés. Remarque : avec le nombre d'individus qui est ici très grand, le dendrogramme est illisible, c'est pourquoi on ne le verra pas.

Voici en Figure 20 le diagramme d'inertie inter-clusters, qui permet de choisir le nombre de groupes optimal pour obtenir la meilleure partition possible de la base de données HYDRoSWOT\_100m\_tronc.

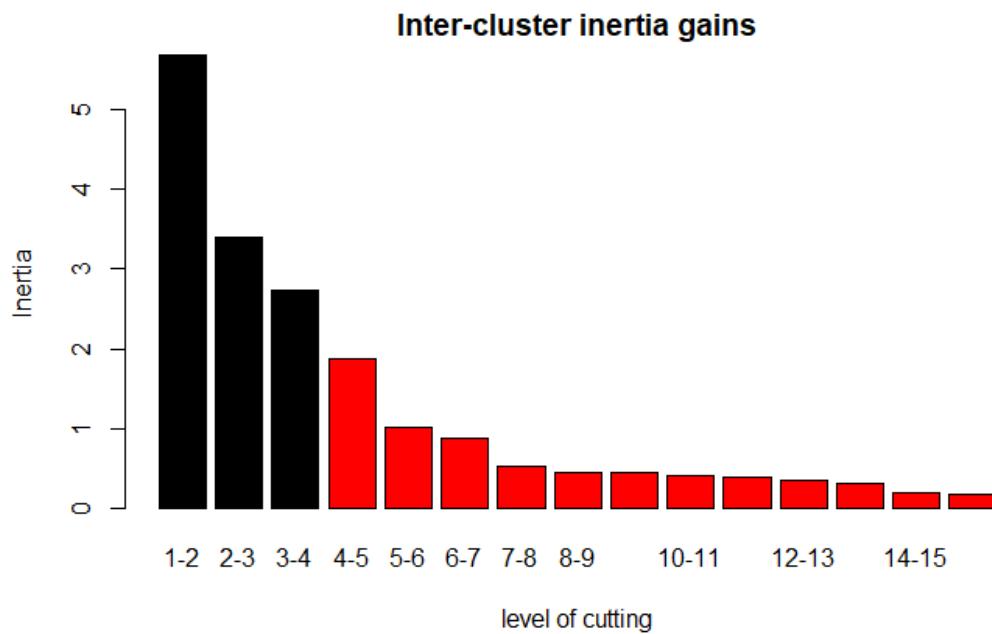


FIGURE 20 – Inter-cluster Inertia - HYDRoSWOT\_100m\_tronc

On constate qu'on couvre la plus grande partie de l'inertie inter-cluster en coupant la base de données en 3 ou 4. On obtient grâce à R le découpage en 4 clusters, dont on trouve la description R dans les tableaux Figure 21 et 22.

A gauche, on a la description des clusters en fonction des rivières les plus représentatives du cluster. A droite, on a la description des clusters par variables le plus significativement associées aux clusters.

Variables the most significantly associated to the cluster		
	Mean in the cluster	Overall mean
silt	59,41481661	48,98532114
LC12	2,971766994	1,448164848
flowacc	70825,6698	35277,00721
dA	708,5376097	292,406595
LC7	76,33863764	57,8695517
W	292,948808	174,206277
dH	2,749996472	1,66475391
Q	1269,472979	594,9001515
LC6	10,25250245	8,735324073
LC8	0	0,006459069
LC3	4,182280398	12,11709315
LC9	1,734819006	8,666113582
LC5	0,021251516	0,537515212
stage	2,896073292	5,989493699
LC2	0,000615988	0,224699355
LC4	3,918227913	8,09339851
LC1	0,578717627	2,270854456
sand	19,2272925	29,51765733

River	Percentage of the cluster
Wabash	99,378882
Iowa	0,62111801
San Joaquin	0
Sacramento Downstream	0
Connecticut	0
Kanawha	0
Ohio	0

(a) Cluster 1

Variables the most significantly associated to the cluster		
	Mean in the cluster	Overall mean
clay	24,25709321	21,47662055
LC6	10,79350667	8,735324073
LC5	0,923450036	0,537515212
stage	8,044806816	5,989493699
LC7	68,14505229	57,8695517
dH	1,505413747	1,66475391
LC4	7,519122716	8,09339851
LC11	0	0,001883886
LC2	0,176828181	0,224699355
sand	28,06842298	29,51765733
LC8	0	0,006459069
LC1	1,935324064	2,270854456
silt	47,64641371	48,98532114
LC9	6,197637033	8,666113582
dA	125,6610836	292,406595
Q	184,5483976	594,9001515
flowacc	19840,40913	35277,00721
LC3	3,679993888	12,11709315
W	89,9088783	174,206277
LC12	0,579636733	1,448164848

(b) Cluster 2

FIGURE 21 – Description of each cluster by river (left) and by variables (right)

Variables the most significantly associated to the cluster		
	Mean in the cluster	Overall mean
LC11	0,115702	0,001883886
LC8	0,322314	0,006459069
flowacc	235814,4538	35277,00721
Q	6004,475327	594,9001515
dA	1912,326114	292,406595
LC9	47,471074	8,666113582
W	649,196314	174,206277
LC12	4,338843	1,448164848
LC1	4,157025	2,270854456
silt	55,505495	48,98532114
dH	2,9845	1,66475391
LC2	0,413223	0,224699355
sand	23,901099	29,51765733
LC5	0	0,537515212
LC6	5,752066	8,735324073
LC7	25,438017	57,8695517

River	Percentage of the cluster
Ohio	100
Wabash	0
Iowa	0
San Joaquin	0
SacramentoDownstream	0
Connecticut	0
Kanawha	0

(a) Cluster 3

Variables the most significantly associated to the cluster		
	Mean in the cluster	Overall mean
LC3	45,82017673	12,11709315
sand	45,59611684	29,51765733
LC1	4,96318009	2,270854456
LC4	14,55582779	8,09339851
LC2	0,596238393	0,224699355
LC9	20,28384169	8,666113582
LC12	2,02695931	1,448164848
W	246,6420159	174,206277
dA	178,1323995	292,406595
flowacc	23815,97657	35277,00721
dH	0,810978207	1,66475391
stage	3,417985159	5,989493699
LC5	0,040011393	0,537515212
silt	40,734365	48,98532114
clay	13,66147226	21,47662055
LC6	1,350185241	8,735324073
LC7	10,3538333	57,8695517

River	Percentage of the cluster
Connecticut	70,3448276
Kanawha	29,6551724
San Joaquin	0
SacramentoDownstream	0
Wabash	0
Iowa	0
Ohio	0

(b) Cluster 4

FIGURE 22 – Description of each cluster by river (left) and by variables (right)

On peut constater que le clustering regroupe les rivières comme on peut le faire avec l'ACP. Le clustering a créé les mêmes groupes que si on avait coupé la projection des individus sur le premier plan factoriel en 4 :

- Le cluster 1 : est constitué presque à 100% de Wabash, la rivière du quart supérieur gauche dans le premier plan factoriel de l'ACP ; soit une rivière à débit élevé, argileuse et limoneuse.

- Le cluster 2 : constitué principalement de Iowa river, Sacramento Downstream, San Joaquin et dans une moindre mesure de Wabash. Ce sont les rivières sur l'axe 1 négatif du premier plan factoriel, soit les rivières argileuses (clay), avec peu d'arbres mais plutôt des herbacées et autour des terres cultivées (LC6, LC7) et un débit Q moyen.
- Le cluster 3 : constitué uniquement de l'Ohio, distincte de toutes les autres rivières par ses valeurs très élevées du débit Q et de l'aire draînée flowacc.
- Le cluster 4, constituée du Connecticut et de Kanawha, les deux rivières sur l'axe 1 positifs, caractérisée par une valeur élevée du taux de sable et d'arbres (LC1, LC2, LC3, LC4).

L'analyse par classification hiérarchique ajoute à l'ACP plus de précision sur les valeurs prises des variables pour les individus d'un même cluster. Sinon, on ne tire pas tellement d'information nouvelle. Les groupes de variables restent identiques. **On apprend simplement que le critère le plus fort pour regrouper des rivières est le type de sol, c'est-à-dire le pourcentage de limon, d'argile ou de sable (silt, clay et sand).**

**Remarque :** Le clustering PEPSI n'apporte aucune information, il ne crée pas de groupe : la base de données est sectionnée en 7, soit les 7 rivières. C'est pourquoi on ne prendra pas la peine de l'étudier.

### 3.2 Analyse de PEPSI\_tronc avec les variables « meandwave » et « sinuosity »

On cherche à savoir quelles informations on peut tirer de la base de données PEPSI\_tronc en ajoutant les variables correspondant à la longueur d'ondes de méandre « meanwave » et à la sinuosité « sinuosity » de la rivière. N'ayant **pas les informations sur la rivière San Joaquin**, la base de données étudiée concerne seulement les six autres rivières des sept rivières étudiées : Ohio, Connecticut, Kanawha, Wabash, Sacramento Downstream et Iowa.

#### 3.2.1 Analyse de la matrice des corrélations des variables

Voici en Figure 23 la matrice des corrélations de la matrice PEPSI\_tronc à laquelle on a ajouté « meandwave » et « sinuosity ». A l'exception des corrélations de ces nouvelles variables avec les anciennes, toutes les autres sont les mêmes que pour la première analyse de corrélations de PEPSI\_tronc. On ne va donc s'intéresser ici qu'aux corrélations des variables avec meandwave et sinuosity. Se référer à la partie 3.1.1. pour les autres corrélations.

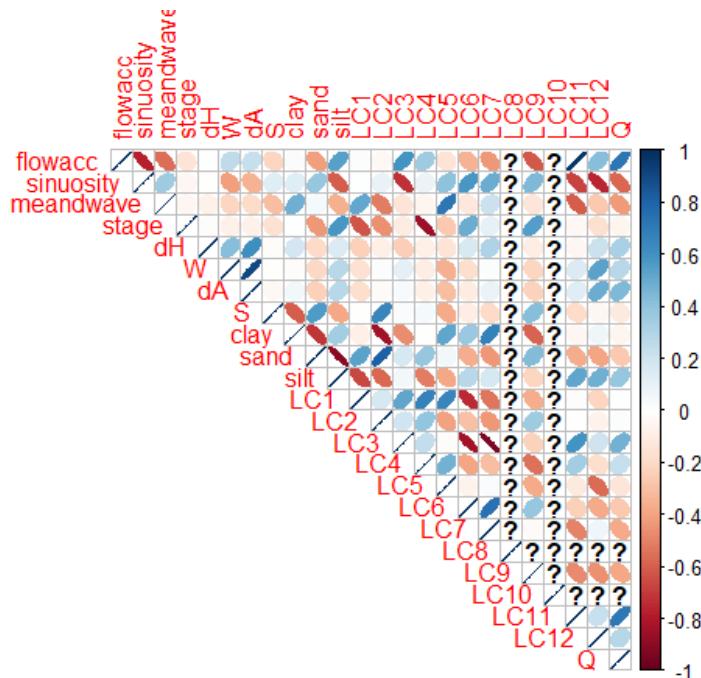


FIGURE 23 – PEPSI database with « meandwave » and « sinuosity » correlation matrix

- Corrélation positive :
  - meandwave et clay, LC1, LC5
  - sinuosity légèrement avec LC5, LC6, LC7 et meandwave
- Corrélation négative :

- meandwave avec silt, LC2 ; LC11 ; LC12, Q, flowacc
- sinuosity avec silt, LC3, LC11, LC12, Q, flowacc

Au vu de ces corrélations, meandwave et sinuosity semblent se mettre de paire à l'opposé du groupe de variable Q, flowacc, W, dA etc. On a un débit plus faible là où la sinuosité est plus forte, et là où la longueur d'onde de méandre est la plus grande. Elles sont légèrement corrélées positivement à LC5, LC6, LC7 et clay. On a donc avec meandwave et sinuosity une influence supplémentaire sur les variables de débit, à la fois légèrement dépendant du groupe de variable d'argile et herbacées.

### 3.2.2 Analyse en Composantes Principales

L'ACP sur PEPSI\_tronc avec les variables meandwave et sinuosity va permettre de nuancer la typologie des rivières. Voici en Figure 24 le diagramme de Pareto pour cette base de données et en Tableau 5 le pourcentage de variance expliquée par composante et la valeur propre associée.

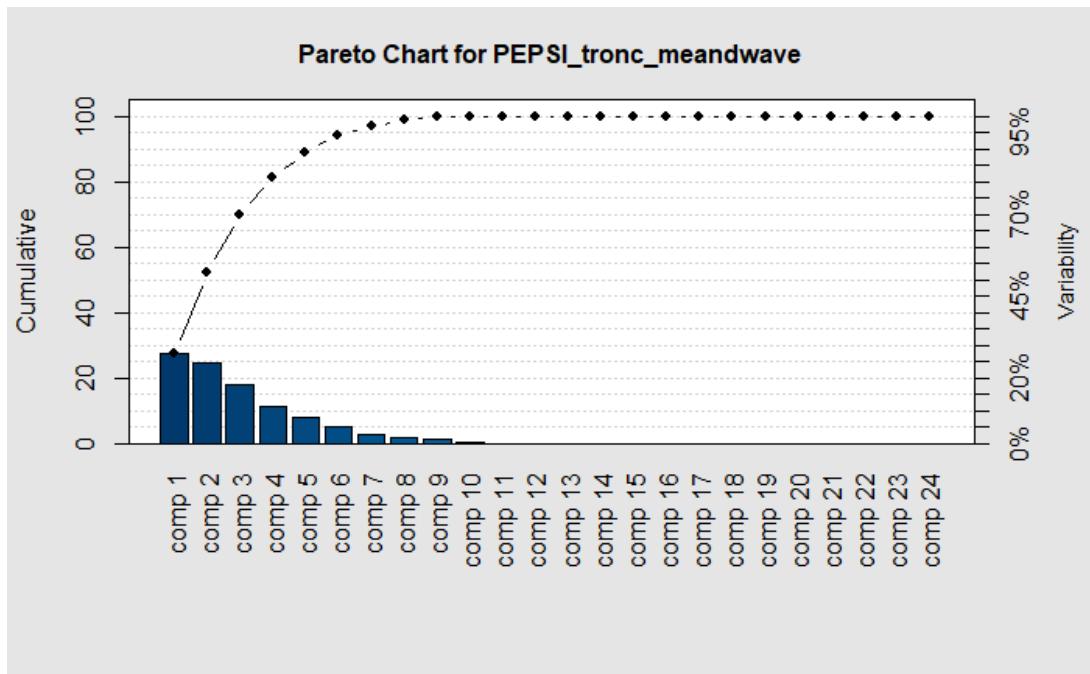


FIGURE 24 – Pareto Chart for PEPSI\_tronc database with « meandwave » and « sinuosity »

	Eigen value	Percentage of variance
comp 1	6.048119	27.491451
comp 2	5.460357	24.819803
comp 3	3.897308	17.715038
comp 4	2.434384	11.065384
comp 5	1.724381	7.838096
comp 6	1.164725	5.294205

TABLE 5 – Percentage of variance per component

Comme pour les deux autres ACP, on ne retient que les deux premières composantes, qui expliquent ici environ 52% de la variabilité entre les individus et l'individu de référence. On peut à présent projeter les variables sur le premier plan factoriel afin d'interpréter les composantes.

Voici en Figure 25 la projection des variables sur le premier plan factoriel, ainsi que le tableau regroupant la contribution précise de chaque variable (1/24) :

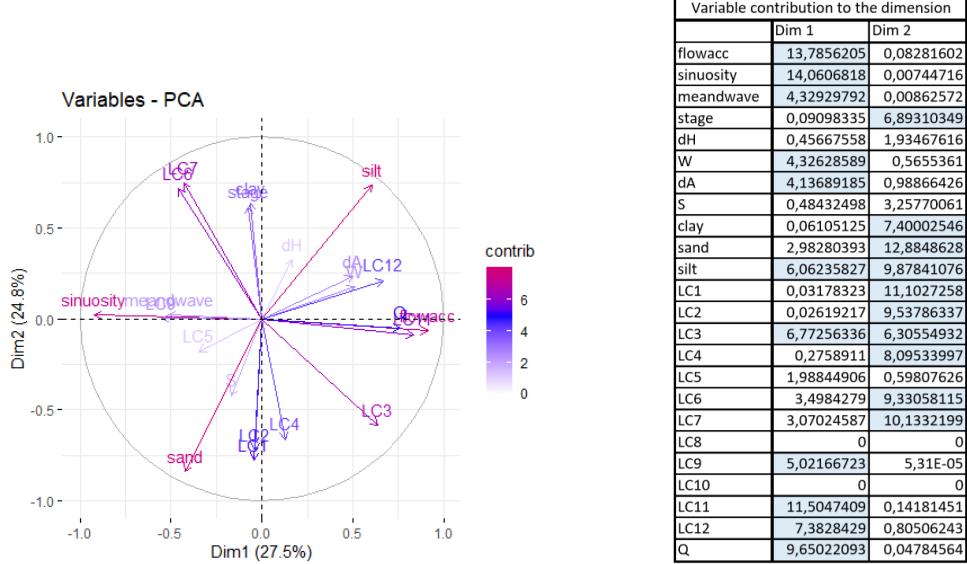


FIGURE 25 – PEPSI\_tronc with meandwave and sinuosity PCA variables plot

Interprétation des composantes : Ici, le nombre de variable étant plus élevé, on retient une variable comme contribuante à partir d'une contribution supérieure ou égale à  $\frac{1}{24}$  c'est à dire environ 4.16.

— Dimension 1 : les variables contribuantes sont, dans l'ordre décroissant :

- pour les positifs : flowacc, LC11, Q, LC12, LC3, silt, W, dA,
- pour les négatifs : sinuosity, LC9, meandwave

— Dimension 2 :

- pour les positifs : LC7, silt, LC6, clay, stage
- pour les négatifs : sand, LC1, LC2, LC4, LC3

On remarque que la composante 1 de l'ACP PEPSI\_tronc\_meandwave est la composante 2 de l'ACP PEPSI\_tronc. En effet, cette composante (constituée des variables sand, clay, LC1, LC2, LC3 etc) explique moins de variabilité entre les individus que l'autre (constituée de Q, flowacc, W, etc), qui est à présent plus nuancé à l'aide des variables « meandwave » et « sinuosity ». Ainsi comme prévu, la composante 1 de cette ACP ci oppose une grande aire draînée et un haut débit à une forte sinuosité et une grande longueur d'onde de méandre.

Voici en Figure 26 la projection des individus par force de représentation ainsi que le biplot.

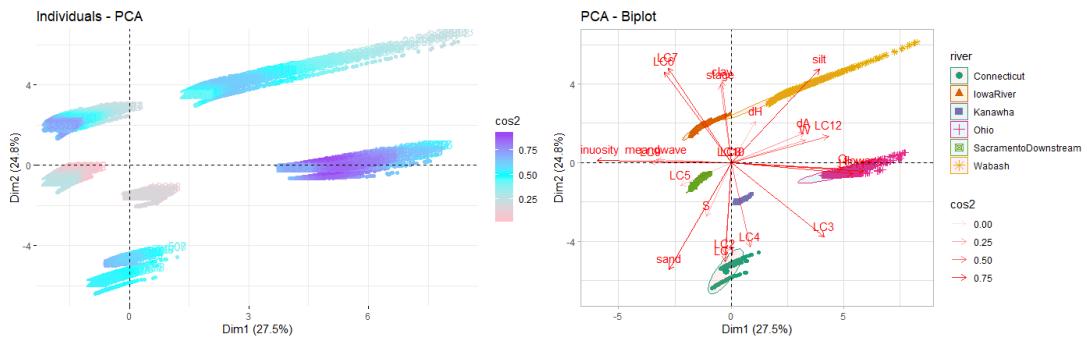


FIGURE 26 – Individual plot by representation power and biplot

On a bien que la répartition des individus sur le premier plan factoriel du diagramme de gauche ressemble à celle de l'ACP de PEPSI\_tronc, mais avec une « rotation » des axes 1 et 2.

- Connecticut : rivière située sur la dimension 2, côté négatif. Elle est caractérisée par une valeur élevée de LC2, LC4, LC1, sand. Ainsi elle est sableuse avec une forte forêt. Même conclusion que pour les autres ACP.
- Kanawha : rivière située dans le quart de plan en bas à gauche. Elle est donc comme le Connecticut, mais avec un débit et une aire draînée un peu plus élevé. Peu bien représenté.
- Ohio : sur les positifs de la dimension 1. Elle est caractérisée principalement par des valeurs élevées de Q, flowacc et faibles de sinuosity, meandwave. Ainsi, c'est une rivière large peu sinuose, avec des méandres peu marqués, un débit élevé et une aire draînée élevée. Très très bien représentée.

- Sacramento Downstream : sur les négatifs de la dimension 1 et 2. Cette rivière a donc des valeurs faibles de Q, flowacc, ainsi que de clay, stage, LC6 et LC7. Elle a par contre des valeurs proche de la moyenne pour sinuosity, meandwave, LC2, LC1, LC4 et sand. On apprend peu de chose et de plus cette rivière est très mal représentée par ces composantes.
- Iowa : située dans le quart supérieur gauche du premier plan factoriel, elle est à des valeurs élevées de clay, stage, LC7, LC6. Elle est donc argileuse et pleine d'herbacée et de terre cultivables lol. Ensuite, son débit et son aire draînée sont un peu en dessous de la moyenne alors que sa sinuosité et longueur d'onde de méandre un peu au dessus.
- Wabash : située dans le quart supérieur droit du premier plan factoriel, elle est caractérisée par des valeurs élevées de silt, Q, flowacc, stage et clay, dA et W. C'est une rivière à débit élevé, assez large, très limoneuse, avec une aire draînée conséquente. Elle a des valeurs faibles de sinuosity, meandwave. Elle est donc peu sinuose et avec de faible méandre.

## 4 Analyse statistique comparative entre les données PEPSI appartenant à différents continents

Dans cette partie on ne s'intéressera qu'aux bases de données PEPSI1 et PEPSI2 combinées, donc à l'intégralité des rivières de PEPSI. On va comparer les corrélations entre les variables des rivières PEPSI par continent.

### 4.1 Géolocalisation des stations PEPSI mondiales

Voici en Figure 27 la carte QGIS des stations américaines PEPSI, en Figure 28 les stations européennes et en Figure 29 les stations asiatiques.

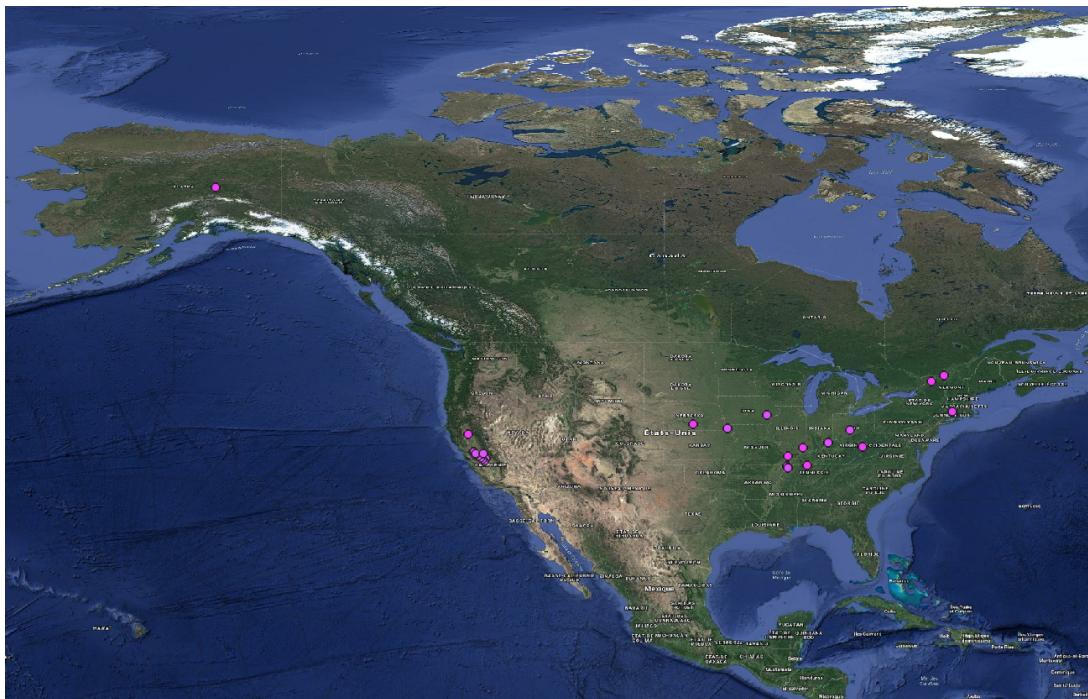


FIGURE 27 – PEPSI american stations - QGIS Satellite map



FIGURE 28 – PEPSI european stations - QGIS Satellite map

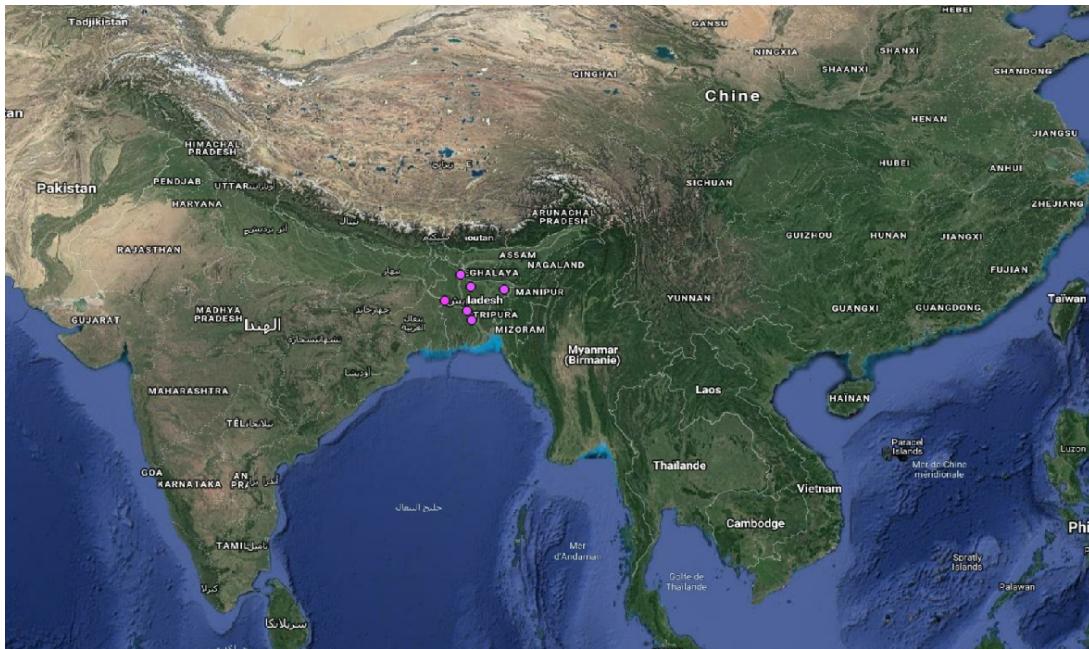


FIGURE 29 – PEPSI asian stations - QGIS Satellite map

On va pouvoir tirer des tendances par continent en analysant les corrélations entre les variables par continent.

## 4.2 Analyse des corrélations par continent

### 4.2.1 Amérique

Voici en Figure 30 la matrice des corrélations entre les variables des rivières américaines PEPSI. La matrice version numérique en Figure 31.

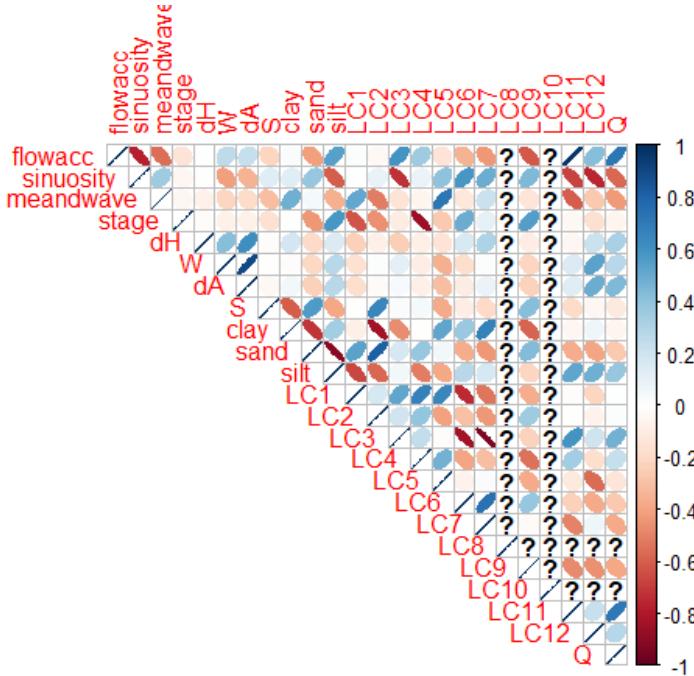


FIGURE 30 – PEPSI american river correlation matrix

Pour l'analyse détaillée des corrélations des variables PEPSI en amérique, voir la partie 3.2.1.

Pour simple rappel, les points clés de cette analyse étaient :

- Le groupe Q, flowacc, W, dH, dA, LC11 et LC12 dont les variables sont corrélées positivement ensemble; opposé à la paire meandwave et sinuosity.
- Le groupe clay, silt, LC5, LC6, LC7 et stage, dont les variables sont corrélées positivement ensemble; opposé au groupe sand, LC1, LC2, LC3, LC4, dont les variables sont corrélées positivement entre elles.
- Le groupe Q, flowacc, W, dH, dA, LC11 et LC12 indépendant des deux groupes cités précédemment.
- La paire meandwave et sinuosity légèrement dépendante du groupe clay, silt, LC5, LC6, LC7.

Pour les comparaisons suivantes, le plus important à retenir est qu'en Amérique du Nord, les rivières ont tendance à avoir un débit élevé et une aire draînée élevée lorsque la sinuosité et la longueur d'onde de méandre sont faibles. De plus, le taux de limon et d'arbre diminuent lorsque la sinuosité et la longueur d'onde de méandre augmentent. L'élévation et la sinuosité sont indépendant entre eux pour les rivières nord-américaine.

	flowacc	sinuosity	meandwave	stage	dH	W	dA	S	clay	sand	silt	LC1	LC2	LC3	LC4	LC5	LC6	LC7	LC9	LC11	LC12	Q
flowacc	1.0000	-0.7744	-0.5536	-0.1446	0.0009	0.2560	0.2303	-0.2151	0.0266	-0.4195	0.5473	0.0151	-0.0394	0.5970	0.3530	-0.1417	-0.3414	-0.4354	-0.6193	0.9716	0.4279	0.7204
sinuosity		1.0000	0.3554	0.0441	-0.0043	-0.4102	-0.3494	0.1358	0.1388	0.3860	-0.6002	0.0044	-0.925	-0.7263	0.0805	0.4062	0.5870	0.4918	0.4416	-0.6732	-0.7515	-0.5719
meandwave			1.0000	-0.0481	-0.0760	-0.2179	-0.1876	-0.3015	0.4804	0.0446	-0.3554	0.5142	-0.5171	-0.1459	-0.0535	0.7277	-0.1251	0.2162	-0.3118	-0.6046	-0.2604	-0.4291
stage				1.0000	-0.0191	-0.0742	-0.0732	-0.1567	0.0089	-0.4394	0.5872	-0.6262	-0.5457	-0.1018	-0.8547	-0.2660	0.4947	0.1026	0.5568	-0.0321	-0.1650	-0.0591
dH					1.0000	0.4294	0.6169	-0.0138	0.1807	-0.1985	0.1534	-0.2331	-0.0820	-0.2446	-0.0836	-0.1461	0.1696	0.3018	-0.0989	-0.0491	0.2165	0.3273
W						1.0000	0.9013	-0.0012	-0.0149	0.2777	-0.1575	0.0147	0.1107	-0.0951	0.3583	-0.1750	0.0158	-0.2123	0.1590	0.5330	0.2707	
dA							1.0000	-0.0359	0.0562	-0.2341	0.2761	-0.1738	-0.0280	0.0395	-0.0838	-0.3080	-0.1099	0.0864	-0.2292	0.1360	0.4959	0.4489
S								1.0000	-0.6097	0.5621	-0.3808	-0.0094	0.6554	0.0218	0.0462	0.3791	-0.0984	-0.1973	0.4287	-0.1871	-0.0452	-0.1159
clay									1.0000	-0.7141	0.3438	-0.0889	0.3887	-0.4635	-0.0050	0.5285	0.3709	0.6757	-0.5869	-0.0481	0.0618	-0.0557
sand										1.0000	-0.9028	0.5329	0.8148	0.1737	0.3890	0.0504	-0.3690	-0.4397	0.4314	0.3657	-0.3974	-0.2627
silt											1.0000	-0.6601	-0.5789	0.0522	-0.5176	-0.3884	0.2702	0.1728	-0.2160	0.5228	0.4881	0.3875
LC1												1.0000	0.1750	0.5206	0.6706	0.6593	-0.7462	-0.5347	0.3685	0.0197	-0.2113	0.0099
LC2													1.0000	0.1974	0.3938	-0.4029	-0.2913	-0.4333	0.3514	-0.0029	-0.0637	0.0121
LC3														1.0000	0.2466	-0.0193	-0.8332	-0.9373	-0.2298	0.5929	0.2021	0.4770
LC4															1.0000	0.4777	-0.3949	-0.3043	-0.5452	0.3412	-0.1765	0.2315
LC5																1.0000	-0.0680	0.0236	-0.3747	-0.1102	-0.5656	-0.1339
LC6																	1.0000	-0.0286	-0.4933	0.0601	-0.3775	
LC7																		1.0000	-0.4727	-0.4589	-0.3848	
LC9																			1.0000	0.2293	0.7129	
LC11																				1.0000	0.2826	
LC12																					1.0000	
Q																						

FIGURE 31 – PEPSI american river numerical correlation matrix

#### 4.2.2 Analyse des corrélations entre les variables PEPSI des rivières d'Europe

Voici en Figure 32 la matrice de corrélations des variables PEPSI des rivières européennes, ainsi que la version numérique en Figure 33.

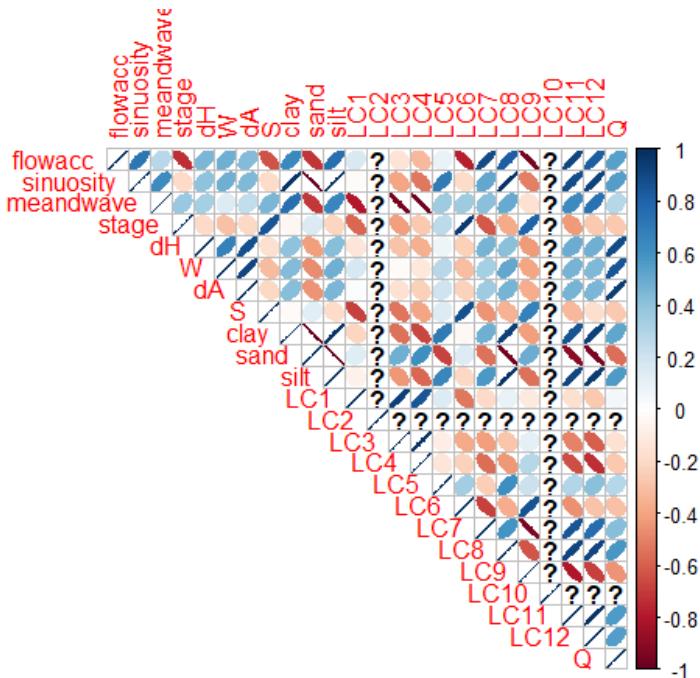


FIGURE 32 – PEPSI european river correlation matrix

— Corrélations positives :

On a plusieurs corrélations qui restent à peu près les mêmes.

— Q, flowacc sont corrélés positivement avec dH, W, dA comme en Amérique du Nord.

— Les variables sinuosity et meandwave sont corrélées positivement entre elles comme en Amérique du Nord. La longueur d'onde de méandre augmente lorsque la sinuosité augmente.

On a des corrélations positives surprenantes :

— Q et flowacc sont ici corrélés positivement avec sinuosity et dans une moindre mesure avec meandwave : l'aire drainée et le débit sont linéairement dépendants de la sinuosité, pour les rivières en Europe et contrairement à celles d'Amérique du Nord. On aura donc que là où les rivières sont plus sinuoseuses en Europe, elles ont aussi une plus grande aire drainée et un plus fort débit, ce qui est plutôt surprenant.

— Les variables sinuosity et clay, silt sont très fortement corrélées. Cela signifie qu'en Europe, le taux d'argile et de limon augmente quand la sinuosité de la rivière augmente. On a aussi que Q et flowacc sont fortement corrélées positivement avec clay et silt. Cela signifie que lorsque les rivières deviennent de plus en plus sinuoseuses, leur débit et leur aire drainée augmentent et en plus de cela le taux d'argile et de limon aussi. Là où les rivières ont un débit élevé, la rivière est aussi boueuse.

— La pente S et l'élévation sont fortement corrélés positivement ensemble : plus on est en altitude, plus la pente de la rivière est forte. En amérique, la pente S n'avait aucune corrélation remarquable.

— Corrélations négatives :

— Les corrélations négatives entre flowacc et S, stage sont très fortes ; plus on est en altitude et plus on est pentu, moins l'aire drainée est grande.

— On a une légère corrélation négative entre stage et sinuosity. La sinuosité augmente quand l'élévation diminue.

— Les variables sinuosity et sand sont très fortement corrélées négativement : le taux de sable diminue lorsque la sinuosité des rivières augmente. Cela est cohérent avec le fait que le taux de limon et le taux d'argile augmentent lorsque la sinuosité augmente. Le taux de sable et le taux d'argile sont systématiquement corrélés négativement.

— La variable meandwave est fortement corrélée négativement avec sand, LC1, LC3, LC4 : le taux de sable, d'arbres diminuent lorsque la longueur d'onde de méandre augmente.

On retient que pour les rivières européennes, contrairement aux rivières américaines, le débit et l'aire drainée augmentent de paire avec la sinuosité, le taux d'argile et de limon. Le débit et l'aire drainée étaient indépendant des variables de qualité du sol pour les rivières américaines. De plus, la longueur d'onde de méandre augmente lorsque le taux de forêt diminue.

	flowacc	sinuosity	meandwave	stage	dH	W	dA	S	clay	sand	silt	LC1	LC3	LC4	LC5	LC6	LC7	LC8	LC9	LC11	LC12	Q
flowacc	1,0000	0,7267	0,2704	-0,7384	0,4550	0,4816	0,4347	-0,6390	0,6426	-0,7149	0,7574	0,1643	-0,1460	-0,3289	0,0905	-0,7774	0,8931	0,8298	-0,9543	0,9134	0,8223	0,5437
sinuosity	1,0000	0,6249	-0,1991	0,4028	0,4830	0,4449	-0,1835	0,9816	-0,9946	0,9974	-0,0299	-0,3905	-0,5280	0,7076	-0,1823	0,5123	0,9821	-0,5000	0,8842	0,9238	0,5594	
meandwa	1,0000	0,3703	0,3305	0,1497	0,2479	0,4589	0,7587	-0,2002	0,6552	-0,2965	-0,9567	-0,9714	0,3551	0,3515	0,4134	0,5077	-0,1593	0,6308	0,7484	0,2801		
stage	1,0000	-0,1918	0,2958	-0,2005	0,8555	-0,0488	0,1511	0,2179	-0,5737	-0,4186	-0,2678	0,2686	0,9527	-0,6121	-0,3714	0,8050	-0,4268	-0,2612	-0,2690			
dH	1,0000	0,6791	0,8758	-0,1607	0,4022	-0,4179	0,4248	-0,1343	-0,2839	-0,3575	0,0728	-0,2387	0,4636	0,4195	-0,4249	0,4978	0,4847	0,8968				
W	1,0000	0,5146	-0,3115	0,4324	-0,6460	0,4822	0,1700	-0,0265	-0,1232	0,2721	-0,3069	0,3359	0,5180	-0,3910	0,4675	0,4447	0,8541					
dA	1,0000	-0,2067	0,4230	0,4425	0,4524	0,0133	-0,1575	-0,2373	0,2087	-0,2280	0,3581	0,4628	-0,3640	0,4618	0,4523	0,9468						
S	1,0000	-0,0219	0,1211	0,1867	0,6878	-0,5374	-0,3961	0,1477	0,8747	-0,4448	-0,3472	0,6712	-0,3239	-0,1703	-0,2625							
clay	1,0000	0,9941	0,9836	0,2113	-0,5483	-0,6623	0,7032	-0,0354	0,4897	0,9334	-0,4199	0,8676	0,9346	0,5251								
sand	1,0000	0,9974	0,9936	0,2113	-0,5483	-0,6623	0,7032	-0,0354	0,4897	0,9334	-0,4199	0,8676	0,9346	0,5251								
silt	1,0000	-0,0755	-0,4349	-0,5745	0,6550	-0,2088	0,5719	0,9822	-0,5469	0,9158	0,9500	0,5660										
LC1	1,0000	0,9303	0,8507	0,1433	-0,5298	-0,1985	0,0965	-0,1177	-0,1604	-0,2694	0,0587											
LC3	1,0000	0,9818	-0,0913	-0,3798	-0,4106	0,2752	0,1165	-0,4912	-0,5983	-0,1587												
LC4	1,0000	-0,1323	-0,2258	-0,5512	-0,4351	0,2836	-0,6471	-0,7353	-0,2609													
LC5	1,0000	0,3496	-0,2438	0,6195	0,2112	0,2966	0,4106	0,2803														
LC6	1,0000	-0,6871	-0,3619	0,8677	-0,4582	-0,2831	-0,2959															
LC7	1,0000	0,5978	-0,9499	0,8530	0,7664	0,4292																
LC8	1,0000	-0,6285	0,9143	0,9185	0,5850																	
LC9	1,0000	-0,8077	-0,6841	-0,4496																		
LC11	1,0000	0,9820	0,5679																			
LC12	1,0000	0,5538																				
Q	1,0000																					

FIGURE 33 – PEPSI european river numerical correlation matrix

#### 4.2.3 Analyse des corrélations entre les variables PEPSI des rivières d'Asie

Voici enfin en Figure 34 la matrice de corrélations entre les variables PEPSI des rivières asiatiques, ainsi que la version numérique en Figure 35.

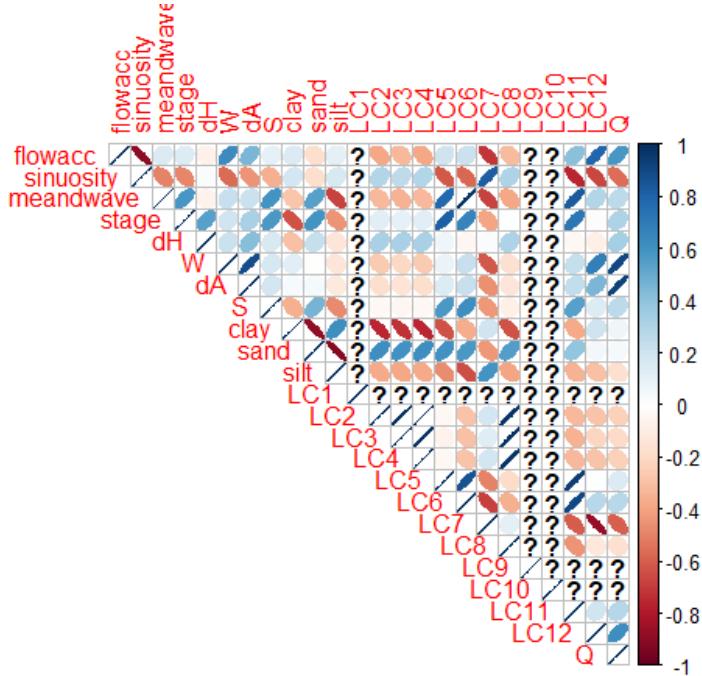


FIGURE 34 – PEPSI asian river correlation matrix

Les rivières d'Asie, autour du Bangladesh dans la base de données PEPSI, ont des points communs avec les rivières nord américaines.

- Corrélations positives :
  - flowacc et Q, W, dA, dH, LC11 et LC12 sont encore une fois corrélées positivement entre eux.
  - clay et silt sont corrélées entre eux
  - sand est corrélée positivement avec LC2, LC3, LC4, LC5.
  - On remarque que LC5 et LC6 sont très fortement corrélées à LC11.
  - La longueur d'onde de méandre meandwave et l'élévation stage sont corrélés positivement.
- Corrélations négatives :
  - Ici flowacc et sinuosity sont fortement corrélées négativement, **comme en Amérique du Nord**. Plus on est sinueux, moins l'aire draînée est grande. De même, sinuosity est corrélée négativement avec Q, LC11 et LC12. Plus on est sinueux, moins le débit est élevé.
  - Cependant, on a étonnamment que sinuosity et meandwave sont corrélées négativement, ce qui est différent de chez les rivières américaines et européennes. On a donc des rivières dont la sinuosité augmente lorsque la longueur d'onde de méandre diminue.
  - Les taux d'argile et de limon, clay et silt sont corrélés négativement avec stage et meandwave. Dans le cas de l'Europe, on avait également cela avec la sinuosité, **cependant ici les corrélations entre sinuosity et clay puis silt sont presque nulles**. On a donc que le taux de limon et d'argile

augmente quand la longueur d'onde de méandre et l'élévation diminue. On ne peut pas dire par contre que ces taux augmentent quand la sinuosité augmente, même si celle ci augmente quand la longueur d'onde de méandre augmente.

- La variable stage et sinuosity sont moyennement corrélés négativement : lorsque l'élévation augmente, la sinuosité de la rivière diminue, comme en Europe. Comme on a que stage et Q sont légèrement corrélés positivement, cela pourrait conduire à dire que les rivières asiatiques ont un débit plus fort en élévation, là où elles sont moins sinuuses.

Les rivières asiatiques semblent avoir un débit plus élevé et une aire draînée plus grande lorsque la sinuosité est faible. Cette conclusion pour les rivières d'Asie est la seule qui est assez marquée par les corrélations pour qu'on puisse l'affirmer sans risques. Les autres liens avec la longueur d'onde de méandre et la sinuosité ne sont pas assez étroits pour pouvoir les considérer comme significatifs. Il serait bien d'avoir plus de données sur les rivières de cette zone d'Asie pour réellement affirmer ou infirmer une grosse partie de l'analyse des corrélations.

	flowacc	sinuosity	meandwave	stage	dH	W	dA	S	clay	sand	silt	LC2	LC3	LC4	LC5	LC6	LC7	LC8	LC11	LC12	Q
flowacc	1,0000	-0,8972	0,1573	0,1401	-0,0790	0,6341	0,4555	0,1081	0,1515	-0,1708	0,1106	-0,3815	-0,3215	-0,3815	0,2061	0,2165	-0,7178	-0,3147	0,4133	0,7933	0,5832
sinuosity	1,0000	-0,4926	-0,4810	0,0256	-0,5621	-0,4409	-0,3543	0,1839	-0,1715	0,1669	0,2961	0,2584	0,2961	-0,6136	-0,5748	0,8279	0,3039	-0,7518	-0,6668	-0,5516	
meandwave	1,0000	0,5885	-0,6685	0,2235	0,2128	0,5939	-0,2748	0,5448	-0,6725	-0,3293	-0,3431	-0,3293	0,7844	0,9878	-0,6714	-0,3815	0,8318	0,2968	0,2680		
stage	1,0000	0,5407	0,1918	0,3289	0,5729	-0,6362	0,5910	-0,4460	0,1284	0,1074	0,1284	0,8105	0,6651	-0,3992	0,0050	0,7267	-0,0124	0,3063			
dH	1,0000	0,2147	0,4232	0,1679	-0,2988	0,2328	-0,1345	0,3292	0,3273	0,3292	0,0666	-0,0415	0,0208	0,3078	-0,0307	-0,0873	0,3363				
W	1,0000	0,8831	0,1750	0,1287	-0,0027	-0,1355	-0,2545	-0,2056	-0,2545	0,0856	0,2263	-0,6137	-0,1653	0,2328	0,6828	0,5066					
dA	1,0000	0,1822	0,0340	0,0409	0,1180	0,1796	-0,1491	0,1796	0,1474	0,2242	-0,4543	0,1343	0,2439	0,4588	0,5282						
S	1,0000	-0,3581	0,4666	-0,4745	-0,0324	-0,0414	-0,0324	0,5700	0,6121	-0,4328	-0,0845	0,5500	0,1491	0,2668							
clay	1,0000	-0,8817	0,6180	0,7500	-0,7298	-0,7500	-0,6319	-0,3627	0,1996	-0,6338	-0,3746	0,2012	0,0673								
sand	1,0000	-0,9150	0,6102	0,6006	0,6102	0,6006	0,6102	0,6006	0,5710	-0,4374	0,5568	0,3922	0,0523	0,0475							
silt	1,0000	-0,3762	-0,3809	-0,3762	-0,4663	-0,6439	0,5933	-0,3914	-0,3427	-0,2992	-0,1636										
LC2	1,0000	0,9966	1,0000	0,0380	-0,2846	0,1866	0,9746	-0,3282	-0,2860	-0,2340											
LC3	1,0000	0,9966	0,0666	-0,2996	0,1379	0,9866	-0,3445	-0,2127	-0,1924												
LC4	1,0000	0,0380	-0,2846	0,1866	0,9746	-0,3282	-0,2860	-0,2340													
LC5	1,0000	0,8678	-0,4912	-0,1900	0,9483	0,0003	0,1561														
LC6	1,0000	-0,6830	-0,3585	0,9017	0,2727	0,2727															
LC7	1,0000	0,1183	-0,6062	-0,8617	-0,5925																
LC8	1,0000	-0,4477	-0,1269	-0,1653																	
LC11	1,0000	0,1925	0,2847																		
LC12	1,0000	0,6152																			
Q	1,0000																				

FIGURE 35 – PEPSI asian river numerical correlation matrix

## Conclusion

L'utilisation des bases de données HYDROSWOT\_100m et PEPSI1, PEPSI2 a mis en lumière le manque de données altimétriques de type SWOT exploitable, problème qui reste à améliorer. En effet, l'exploitation de la base de données HYDROSWOT\_100m a nécessité beaucoup de manipulation préalable sur la base pour être possible, et de même pour sa comparaison avec les bases PEPSI. Il est important de retenir que l'échelle géographique des données est le critère le plus vital de cohérence, et le seul qui peut permettre une analyse significative.

Les variables d'intérêt pour expliquer la variabilité entre les rivières du débit sont les variables flowacc, dA, W, sinuosity, meandwave, et silt. Les autres variables permettent de dresser une typologie des rivières nord-américaine, mais n'influent pas l'évolution des variables de débit. On retrouve des tendances différentes, assez intéressantes par continents, notamment entre meandwave et sinuosity, ou entre flowacc et meandwave/sinuosity.