Chapitre 8 : Traitement des données

Dans la définition des normes pour les traitements de données, il est important de distinguer la résolution, la précision, les erreurs et l'incertitude **[27]**. Cette tâche doit être accomplie en relation avec les normes internationales qui figure dans le « Guide des pratiques hydrologiques » et le « Guide des pratiques climatologiques » publié par l'OMM, qui ensuite, traitée automatiquement par les progiciels « RClimdex », « RHtests_dlyPrcp » et « RHtestsV4 ».

Effectivement, l'organisation des données et les traitements préliminaires ont été préalablement faits par les Ingénieurs, les Techniciens et les observateurs de la DGM. Donc, nos données initiales sont à l'état brute et notre tâche se situe au niveau du contrôle et de la validation des données de précipitations. Mais avant tout, nous manipulerons les bases de données sous Excel pour obtenir les dispositions appropriées à chaque simulation. Toutes ces procédures sont très importantes pour éviter que les résultats ne soient pas erronés. Le chapitre se divise comme suit :

- Manipulation des bases données avec Excel
- Contrôle de Qualité avec Rclimdex
- Correction des erreurs connues
- Validation des données avec RHtests_dlyPrcp et RHtestV4

Section 1 Manipulation des bases de données avec Excel

En premier lieu, les données issues des différents établissements sont tous sous format, soient « .txt », soient « .dat ». Les logiciels utilisés requièrent certaines spécifications sur les formats et les dispositions des données dans les fichiers. Et vue l'importance de la quantité de données à traiter cela demanderait un temps relativement long pour les traitements manuels, mieux vaut choisir la procédure automatique. Pour se faire, Excel sera employé en vue de faire toutes les modifications possibles des bases de données. Pour cela, plusieurs fonctions d'Excel seront employées comme les fonctions bases de données, les tableaux croisées dynamiques, les requêtes de classeurs et de feuilles, les transpositions, les fonctions statistiques de bases et bien d'autres qui dépendent de la disposition voulue. Après modification avec Excel, les fichiers seront reconvertis sous leur format initial pour que les logiciels puissent les lire correctement.

Section 2 Contrôle de qualité sous RClimdex

Dans le cadre du contrôle de qualité, l'objectif est que l'ensemble des données atteigne le meilleur niveau possible avant leur utilisation. Le RClimDex QC effectue les procédures suivantes :

I. Codage des valeurs manquantes

Communément connues comme lacune dans les séries pluviométrique. Ici, le progiciel remplace toutes les valeurs manquantes (actuellement codées comme -99,9) dans un format interne que R reconnaît, c'est-à-dire NA (Not Available), puis produit un schéma sur tout le jeu de donnée.

II. Les valeurs négatives

R va identifier toutes les valeurs déraisonnables dans NA. Ces valeurs sont les valeurs précipitations quotidiennes qui s'élève à moins de zéro (négatives)

III. Détection des valeurs aberrantes

Ce sont des valeurs quotidiennes en dehors d'une région définie par l'utilisateur. Actuellement, cette région est définie comme la moyenne plus ou moins n fois l'écart type de la valeur pour la journée, ce qui est, [moyenne - n * std, signifie + n * std]. StD (Standard Deviation) représente l'écart type de la journée et n est une entrée de l'utilisateur et la moyenne est calculée à partir de la climatologie de la journée (dans notre simulation, nous utiliserons n=4)

IV. Simulation sous RClimdex

Le contrôle de qualité des données se fait à partir de la fonction ***Run QC*** de l'interface graphique GUI du progiciel RClimdex. Mais avant, on charge les données sous forme de format Texte par la fonction ***Load data***. La procédure sera présentée avec l'interface GUI dans l'Annexe X.

Section 3 Correction des erreurs connues

Nous présenterons ici les procédures implantées dans RClimdex.

I. Elimination des lacunes de longue période

Dans cette section, nous allons procéder à l'élimination des mois et des années contenant respectivement plus de 10 jours et 3 mois de manques selon les normes de l'OMM.

II. Analyse des valeurs aberrantes

Ici, nous examinerons les données aberrantes, si elles ont une quelconque relation avec l'état du climat au moment de l'observation. Si oui, nous ne les modifieront pas. Si non, on les éliminera.

III. Restitution des valeurs manquantes

Cette étape sera envisagée si les périodes ne sont pas trop longues, conforme à ce qui a été mentionné dans la section 3.I. Le traitement des données manquantes est très compliqué, il dépend de leur nature et des traitements statistiques que nous souhaitons réaliser par la suite (ex. un remplacement n'a pas le même impact selon que l'on fait une ACP ou une régression par la suite) **[77]**. Pour traiter les valeurs manquantes, nous allons suivre trois étapes : la répartition des valeurs manquantes, la classification et les méthodes de traitement. Mais avant cela, on va présenter les critères établis par l'OMM concernant les séries des précipitations manquantes. Les critères sur les données manquantes sont expliqués dans l'Annexe XI.

1. Répartition des valeurs manquantes

La répartition des données manquantes révèle leurs natures. On distingue généralement trois types de « pattern » pour les données manquantes (Figure 22).

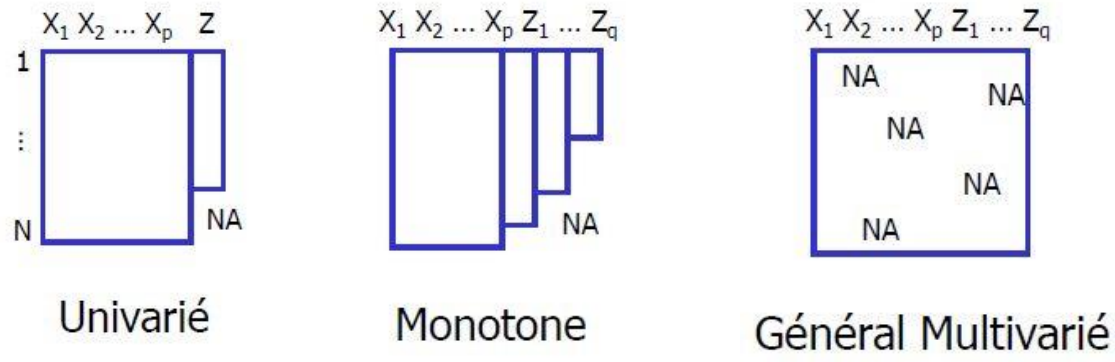

Figure 22 : Schéma de la répartition des données manquantes
Source : **[75]**

2. Classification des données manquantes

[50] Dans la littérature, il existe trois hypothèses distinctes sur l'origine du mécanisme des données manquantes :

❖ **MCAR (Missing Completely At Random)** : les données sont manquantes complètement aléatoirement si la probabilité d'avoir une valeur manquante pour une variable donnée ne dépend pas de celle-ci, mais uniquement des paramètres extérieurs indépendants de cette variable.

❖ **MAR (Missing At Random)** : les données sont manquantes aléatoirement si la probabilité d'avoir une valeur manquante peut dépendre des observations mais pas des données manquantes.

❖ **MNAR (Missing Not At Random)** : les données sont manquantes non aléatoires lorsque la probabilité de non-réponse est liée aux valeurs prises par la variable ayant des données manquantes.

Dans le cas des données de précipitations les données manquantes sont de classe MCAR puisque les séries manquantes ne dépend pas d'eux-mêmes, mais uniquement des paramètres extérieurs indépendants, notamment les conséquences des erreurs aléatoires et systématiques.

IV. Méthode de traitement

Il existe de très nombreuses méthodes pour le traitement des données manquantes, elles consistent à substituer à la valeur manquante une valeur choisie de manière « pertinente ». Cependant, il faut faire attention puisque les méthodes sont « séduisantes et dangereuses » (D.Dubin), on doit bien maîtriser la nature de tous le jeu de données complet. Tous ces méthodes sont regroupées en deux grandes catégories **[49]** :

- Imputation Simple
- Imputation Multiple

Notons toujours que nous appliquons les normes utilisées par l'OMM dans l'intégralité des méthodes de traitement des données.

1. Imputation Simple (IS)

Méthode typique d'un processus d'observation MAR, or nos données de précipitations sont de classes MCAR. Mais conformément aux normes, l'IS se réfère à l'estimation des données manquantes à partir des valeurs provenant des stations voisines soumises aux mêmes conditions climatiques et situées dans la même zone géographique. Trois méthodes sont proposées pour les données pluviométriques **[11]** :

- Remplacer la valeur manquante par celle de la station la plus proche
- Remplacer la valeur manquante par la moyenne des stations voisines
- Remplacer la valeur manquante par une moyenne pondérée par la tendance annuelle des stations pluviométriques, soit :

$$P_x = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\overline{P_x}}{\overline{P_i}}P_i\right) \qquad (11)$$

Où :

P_x : donnée manquante de précipitation estimée

n : nombre de stations de référence,

P_i : précipitation à la station de référence i,

$\overline{P_x}$: précipitation moyenne à long terme de la station x,

$\overline{P_i}$: : précipitation moyenne à long terme de la station de référence i.

Puisque nous n'avons pas à notre disposition d'autres station de référence, cause de la difficulté d'obtention des données. Les méthodes se référant à l'IS ne sont pas envisageables. De plus, l'inconvénient de cette approche est qu'elle conduit à une sous-estimation parfois violente de la variance des estimateurs, en ajoutant aussi l'importance du temps de traitement qui est contraire aux résultats attendus.

2. Imputation Multiple (IM)

[74] [75] [41] [49] Cette méthode consiste à créer plusieurs valeurs possibles d'une valeur manquante (plusieurs imputations différentes en se basant sur l'ensemble des données disponibles). On analyse la matrice des données pour en déduire un modèle pour les valeurs manquantes en tenant compte de cette multiplicité, puis on réalise entre M = 3 et 10 imputations pour obtenir des jeux de données correspondants. Ensuite, on calcule le paramètre d'intérêt pour chaque jeu pour enfin combiner les M imputations en vue d'obtenir une inférence qui tienne compte de l'incertitude supplémentaire liée aux valeurs manquantes. Les objectifs sont :

- De refléter correctement l'incertitude des données manquantes
- De préserver les aspects importants des distributions
- De préserver les relations importantes entre les variables

Les buts ne sont pas de prédire les données avec la plus grande précision ni de les décrire de la meilleure façon possible, mais de minimiser l'incertitude entre la réalité et l'estimation.

En météorologie et climatologie, on distingue deux types de méthodes intégrant l'IM :

- L'Algorithme EM (Expectation-Maximisation)
- Les méthodes de calcul MCMC (Monte Carlo Markov Chain)

Il s'avère que ces méthodes sont pratiquement impossibles sans traitement automatique par l'intermédiaire des logiciels statistiques puissants et robustes tels que R, S+, SPSS, SAS, WinBUGS. Dans notre cas, nous utiliserons SPSS. Pour le choix des méthodes, on va opter pour l'IM à l'aide de l'Algorithme EM en vue de sa souplesse et de sa considération sur toutes l'ensemble des données disponibles. Contrairement à la chaîne de Markov qui ne prend en compte que les données du temps présent. De plus, la simulation est rapide en vue des résultats avec le minimum d'incertitude.

L'Algorithme EM (Expectation-Maximisation)

L'algorithme EM ou encore Expectation-Maximisation est un algorithme itératif conçu par Dempster, Laird et Rubin (1977). Il s'agit d'une méthode d'estimation paramétrique s'inscrivant dans le cadre général du maximum de vraisemblance.

❖ **Méthode d'estimation des paramètres** : Il existe plusieurs méthodes pour estimer les paramètres d 'une distribution statistique. La méthode du Maximum de vraisemblance est appropriée pour l'estimation des paramètres de l'algorithme EM.

[21] Cette méthode consiste, pour un échantillon donné, à maximiser la fonction de vraisemblance (fonction de densité jointe) par rapport aux paramètres. Cette méthode est habituellement complexe (le logiciel le traitera pour nous) mais a généralement des propriétés asymptotiques intéressantes.

Soit un échantillon aléatoire (une partie de la série de précipitations) $X_1, X_2, \dots, X_n$ tiré d'une distribution $F(x; \theta_1, \theta_2, \dots, \theta_p)$. Lorsqu'ils existent, les estimateurs obtenus par la méthode du maximum de Vraisemblance sont les solutions $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p$ du système de p équations :

$$\frac{\partial \mathcal{L}(\theta_1, \theta_2, \dots, \theta_p)}{\partial \theta_t} = 0 \qquad r = 1,2, \dots, p \qquad (12)$$

Ou la fonction de vraisemblance est définie par :

$$\mathcal{L}(\theta_1, \theta_2, \dots, \theta_p) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \dots, \theta_p) \qquad (13)$$

❖ **<u>Principe</u>** : **[75] [50]** On analyse l'interdépendance entre les paramètres θ (issues de la fonction de vraisemblance) et Y_{mqt} (données manquantes). Y_{mqt} contient de l'information utile pour estimer θ qui permet d'obtenir des valeurs pertinentes pour les données manquantes. On remplit les valeurs manquantes à partir d'une estimation de θ puis on réestime θ à partir de Y_{mqt} et Y_{obs} (données observées), enfin on répète jusqu'à convergence.

Les données de précipitations complètes (i.e., Y_{mqt} et Y_{obs}) peuvent être mise sous la forme suivante :

$$Pr(Y|\,\theta) = Pr(Y_{obs}|\theta)Pr\big(Y_{mqt}|Y_{obs},\theta\big) \qquad (14)$$

D'où :

$$\mathcal{L}(\theta|Y) = \mathcal{L}(\theta|Y_{obs}) + \log Pr\big(Y_{mqt}|Y_{obs},\theta\big) + c \qquad (15)$$

Avec $\mathcal{L}(\theta|Y) = \log Pr(Y|\,\theta)$ vraisemblance des données complètes, et $\mathcal{L}(\theta|Y_{obs}) = \log Pr(Y_{obs}|\theta)$ vraisemblance des données observées. $Pr\big(Y_{mqt}|Y_{obs},\theta\big)$ est la distribution prédictive des données manquantes sachant θ et fait le lien entre les deux.

L'algorithme EM est une procédure itérative en 2 étapes. Soit $\theta^{(t)}$ l'estimation courante de θ, les estimations se font ensuite en 2 étapes :

❖ **E (Expectation)** : étape qui donne la log-vraisemblance

$$\mathcal{Q}\big(\theta|\theta^{(t)}\big) = \int \mathcal{L}(\theta|Y)\ Pr\big(Y_{mqt}|Y_{obs},\theta = \theta^{(t)}\big)dY_{mqt} \qquad (16)$$

❖ **M (Maximisation)** : étape qui détermine $\theta^{(t+1)}$ en maximisant cette log-vraisemblance

$$\mathcal{Q}\big(\theta^{(t+1)}|\theta^{(t)}\big) \geq \mathcal{Q}\big(\theta|\theta^{(t)}\big) \qquad (17)$$

Un résultat de Dempster, Laird et D. Rubin (1977) montre que si $\theta^{(t+1)}$ est la valeur de θ qui maximise $\mathcal{Q}\big(\theta|\theta^{(t)}\big)$ alors $\theta^{(t+1)}$ est une meilleure estimation que $\theta^{(t)}$ car la vraisemblance des données observées pour $\theta^{(t+1)}$ est au moins aussi grande que celle pour $\theta^{(t)}$. Ceci se traduit par :

$$\mathcal{L}\big(\theta^{(t+1)}|Y_{obs}\big) \geq \mathcal{L}\big(\theta^{(t)}|Y_{obs}\big) \qquad (18)$$

Cette méthode sera traitée avec le logiciel IBM SPSS Statistics 24, dans le menu ***Analyse***, la sélection ***Imputation multiple***, puis ***Imputer les valeurs des données manquantes…***. La procédure qu'on va suivre est la suivante, la procédure avec l'interface SPSS sera dans l'Annexe XII :

- Première étape : On restitue d'abord les valeurs manquantes journalières de précipitations en éliminant les mois contenant 10 jours et plus de manques.
- Deuxième étape : Ensuite, la restitution des valeurs mensuelles en éliminant les années contenant 3 mois et plus de manques.
- Troisième étape : Enfin, la dernière étape est la restitution des années manquantes.

Section 4 Validation des données

Après le contrôle de qualité, on va procéder aux validations des données de précipitations. Comme les précipitations constituent un phénomène hydrologique très important et particulièrement variable, elles nécessitent des procédures de validation particulières. Plusieurs méthodes sont proposées suivant les normes de l'OMM. Ici, Un progiciel pour l'utilisation de l'algorithme transPMFred pour détecter les changements dans les quantités de précipitations journalières non nulles a été développé et mis à disposition en ligne, ainsi qu'un algorithme d'équivalence quantique (QM) pour ajuster les quarts de travail dans des séries de précipitations quotidiennes non nulles, applicable à tous les résultats positifs des données.

I. Homogénéisation des données avec RHtests_dlyPrcp et RHtestsV4

Des inhomogénéités ou des « sauts », ou des « ruptures » soudains dans les données peuvent se produire dans des séries de données climatiques soit en raison d'un véritable changement progressif du climat, soit à la suite de changements dans la manière dont les données sont enregistrées (normalement, une Modification de l'emplacement de la station ou de l'équipement d'enregistrement utilisé). L'objectif de l'homogénéisation des séries de précipitations est de détecter et d'identifier ces sauts, ensuite d'ajuster la série par différentes méthodes statistiques. Vue que nous n'avons pas de série de référence, les méthodes proposées par le progiciel RHtests_dlyPrcp sont les plus appropriées. RHtests_dlyPrcp et RhtestsV4 utilisent une technique de régression en deux phases (Wang 2003) pour la détection et l'ajustement de l'inhomogénéité **[59] [31]**.

1. Détection et identification des points de changement

La méthode utilisée est celle proposée par Wang (2008a et 2008b) dans le progiciel RHtests_dlyPrcp appelée « TransPenalized Maximal F-test (transPMFred) ». Cette méthode intègre une procédure de transformation de puissance de Box-Cox dans un test de régression en deux phases à tendance commune (la version étendue du test F maximal pénalisé ou l'algorithme " PMFred ") pour détecter les points de changement pour effectuer le test applicable aux séries de données non gaussiennes, telles que les quantités de précipitations journalières non nulles ou la vitesse du vent. Les aspects de détection-puissance de la méthode transformée (transPMFred) sont évalués par une étude de simulation qui montre que ce nouvel algorithme est beaucoup mieux que la méthode non transformée correspondante pour les données non-gaussiennes ; la procédure de transformation peut augmenter le taux de réussite jusqu'à 70%. Les points de changement détectés sont en bon accord avec des temps de modification documentés pour toutes les séries d'exemples. Cette méthode clarifie qu'il est essentiel pour l'homogénéisation des séries quotidiennes de données sur les précipitations de tester les séries de quantité de précipitations non nulles et la série de fréquences d'occurrence de précipitation (ou absence de courant), séparément. Le nouveau transPMFred peut être utilisé pour tester la série de précipitations

journalières non nulles (qui ne sont pas gaussiennes et positives), et l'algorithme PMFred existant peut être utilisé pour tester la série de fréquences.

La Transformation de la puissance Box-Cox (Box et Cox 1964) est nécessaire, car les quantités quotidiennes de précipitations ne sont normalement pas distribuées. Comme les précipitations quotidiennes sont très variables à la fois spatialement et temporellement (il pourrait pleuvoir dans ce côté de la rue, mais pas l'autre côté), il est difficile de trouver une série de référence appropriée (sauf dans le cas de mesures parallèles).

Dans le cas des séries mensuelles et annuelles qui suivent approximativement une loi normale, l'algorithme PMFred suffit sans la Transformation de la puissance Box-Cox sous RhtestsV4 **[59]**.

2. Ajustement de la série

[54] La méthode proposée par Wang et al. (2009) est l'ajustement « QM Quantile-Match ». L'algorithme QM a également été développé pour ajuster les données gaussiennes telles que nos séries de précipitations quotidiennes. On remarque que les discontinuités de fréquence sont souvent inévitables en raison des changements dans la précision de mesure des précipitations et qu'elles pourraient compliquer la détection de décalages dans des séries de données quotidiennes de précipitations non nulles et annuler toute tentative d'homogénéisation de la série. Dans ce cas, il faut tenir compte de toutes les discontinuités de fréquence avant de tenter d'ajuster les quantités mesurées. Cette méthode propose également des approches pour tenir compte des discontinuités de fréquence détectées, par exemple, pour compléter les mesures manquées des petites précipitations ou les rapports manqués de précipitations traces. Il souligne l'importance de tester l'homogénéité de la série de fréquence des précipitations nulles rapportées et de divers petits événements de précipitation, ainsi que le test des séries de quantités quotidiennes de précipitations supérieures à une petite valeur seuil, en faisant varier le seuil sur un ensemble de petites valeurs qui reflètent les changements dans la mesure de la précision au fil du temps.

II. Simulation sous RHtests_dlyPrcp et RHtestV4

Pour la détection et l'identifications des ruptures, ainsi que l'ajustement QM ; on utilise les fonctions ***FindU*** et ***FindUD*** de l'interface graphique GUI. On se servira de la fonction ***StepSize*** que l'on répétera en boucle de façon à ce qu'on ait un changement significatif, pour réévaluer la signification et l'ampleur des points de changement. La procédure avec l'interface GUI sera présentée dans l'Annexe VII et VIII.

Section 5 Synthèse du traitement des données

Pour conclure ce chapitre, en partant du contrôle de qualité jusqu'aux validations, les analyses et calculs ont été faites automatiquement l'aide des logiciels. Nous n'avons fait qu'expliquer le fonctionnement du traitement dans les logiciels. Maintenant, nous pouvons caractériser le climat à l'aide de jeux de données avec le minimum d'incertitude et le maximum d'efficacité.

Chapitre 9 : Statistique descriptive

La procédure Descriptive permet de décrire la distribution d'une variable continue d'intervalle ou de rapport. Les mesures de tendances centrales et de dispersion constituent la base sur laquelle s'appuient les analyses descriptives pour ce type de variable. Ici, nous utiliserons les modèles proposés par IBM SPSS Statistics 24. On utilise les fonctions ***Fréquences***, ***Descriptive*** et ***Explorer*** (Annexe XVIII). L'objectif de cette méthode est :

- De dégager les propriétés essentielles que l'on peut déduire de l'accumulation des données de précipitations.
- De donner une image concise et simplifiée de la réalité de distributions des séries de précipitations par rapport à l'évènement extrême de 2016-2017.
- Décrire la nature de l'évènement extrême de l'année 2016-2017.

Précisons aussi que les hauteurs de pluies sont des variables continues (peut prendre n'importe quelle valeur sur un intervalle défini). De ce fait, notre série de précipitation est un ensemble de variables aléatoires réelles continues noté $\{X\}$. Les propriétés de statistiques descriptives que nous utiliserons sont les suivantes :

Section 1 Les Quantiles

On les appelle aussi « classes » en statistiques élémentaires continues. C'est une propriété propre des variables aléatoires continues **[2]**, comme ce qui est le cas pour les hauteurs de précipitations. Les quantiles sont des caractéristiques de position partageant la série statistique ordonnée en k parties égales pour aboutir à une meilleure vision qualitative des données. Il y a donc un quantile de moins que le nombre de groupes créés.

L'idée est de rapporter quelques points le long de la distribution cumulative des fréquences. Avec ces quelques points, le lecteur peut extrapoler pour obtenir la distribution complète. Le nombre de points à rapporter est variable, mais souvent, on utilise les quartiles (N= 4 points), les déciles (N= 10 points) et les centiles (N= 100 points). Bien entendu, le nombre de points rapportés N doit être nettement inférieur au nombre d'observations n, pour que les valeurs des quantiles soient stables.

[2] En général, le nombre de classes est compris entre 5 et 20 ; il dépend du nombre N d'observations et de l'étalement des données. La formule de Sturges donne une valeur approximative du nombre k de classes :

$$k \cong 1 + 3{,}222 \log_{10} N \qquad (19)$$

Et l'amplitude des classes E/k ou $E = x_{max} - x_{min}$

I. Les Quintiles

Pour notre cas, Les quintiles des précipitations servent à établir une relation entre une hauteur totale mensuelle de précipitation observée et la distribution de fréquence des valeurs observées au cours

de la période pour laquelle les normales ont été calculées sur une période de 30 ans. On classe les 30 valeurs mensuelles ou annuelles par ordre croissant, puis on les divise en 6 quintiles, le premier et le dernier quintile constituent les valeurs extrêmes de la série étudiée **[27]**.

II. Les Quartiles

L'utilisation des quartiles a pour but principal d'observer la variabilité et les valeurs extrêmes des précipitations sur une période. On va l'appliquer sur les graphiques de boîte à dispersion ou Box-Plot.

Un quartile est chacune des trois valeurs qui divisent les données triées en quatre parts égales, de sorte que chaque partie représente 1/4 de l'échantillon de population. **[2]** De ce fait, les quartiles sont 3 nombres Q1, Q2, Q3 tels que :

- 25% des valeurs prises par la série sont inférieures à Q1.
- 25% des valeurs prises par la série sont supérieures à Q3.
- Q2 est la médiane Me.
- Q3-Q1 est l'intervalle interquartile, il contient 50% des valeurs de la série.

Section 2 Les distributions de fréquences

Il s'agit d'un diagramme ou tableau montrant avec quelle fréquence chaque valeur ou chaque série de valeurs d'une variable apparaît dans un ensemble de données. Il est aussi à noter que l'analyse conditionnelle des fréquences est particulièrement utile pour élaborer des scénarios climatiques et pour déterminer les incidences locales de phénomènes tels que le phénomène El Niño-Oscillation australe (ENSO) et IOD pour notre cas, ainsi que d'autres mécanismes de téléconnections **[27]**. On distingue la :

I. Fréquence absolue ou effectif :

[2] Notée n_i, Elle est associée à une valeur x_i de la variable aléatoire $\{X\}$, le nombre d'apparition de cette variable dans la population ou dans l'échantillon.

II. Fréquence relative :

Elle est associée à la valeur x_i de la variable aléatoire $\{X\}$, le nombre : $f_i = \frac{n_i}{n}$, ou n le nombre total de données.

III. Fréquence cumulée absolue :

Elle est associée à une valeur x_i de la variable, le nombre d'individus dont la mesure est inférieure ou égale à x_i. Notée : $N_i = \sum_{k=1}^{i} n_k$

IV. Fréquence cumulée relative :

Elle est définie par, $F_i = \sum_{k=1}^{i} f_k$