Laboratoire des Sciences du Climat
et de l'Environnement

Upscaling et Downscaling dans la modélisation hydrologique

Auteur

Mathis Deronzier
Mines Saint-Étienne

Maîtres de stage

Emmanuel Mouche
C.E.A.
Mathieu Vrac
C.E.A.

Stage de recherche de master 2

Avril-Septembre
2021

Table des matières

1 Downscaling

Le downscaling est une méthode statistique utilisé dans les sciences du climat permettant d'améliorer les modèles de prédiction. À partir des données obtenues par un simulateur S (lui-même reposant sur un modèle physique de prédiction climatique ex : température, pression, évapo-transpiration...) et des données réelles on cherche à corriger les biais systématiques. Le nom "downscaling" vient du fait que l'on cherche souvent à faire des prédictions sur un point particulier du domaine prédit par le simulateur. Cette méthode est très utile dans la pratique où l'on a des simulateurs donnant des informations sur des maillage de grande distance de grille (de l'ordre d'une centaine de kilomètres). Dans notre cas, nous utilisons le downscaling pour prédire les variables de *précipitation* et d'*évapotranspiration* sur le bassin du **Little Washita**.

Pour formuler rigoureusement l'approche du downscaling nous introduisons des hypothèses communément admises dans les sciences du climat. On suppose que les variables étudiées sont des variables aléatoires dépendantes du temps et de l'espace.

Définition 1. *Pour une variable quantitative V à valeur dans $\mathbb{R}$, on appelle $\mathcal{T}_V$ la fonction donnant les valeurs réelle de cette variable sur la terre à un moment donné, formellement (en considérant la terre comme une sphère $\mathcal{S}(\mathbb{R}^3)$) nous avons*

$$\mathcal{T}_V : \mathbb{R}_+ \times \mathcal{S}(\mathbb{R}^3) \quad \to \quad \mathbb{R}. \tag{1}$$

Alors, $\mathcal{T}_V(t,x)$ est la valeur de la variable au temps t au point de coordonnée x sur terre.

Définition 2. *On appelle simulateur de variable quantitative V à valeur dans $\mathbb{R}$, une fonction S_V satisfaisant :*

$$S_V : \mathbb{R}_+ \times \mathcal{S}(\mathbb{R}^3) \quad \to \quad \mathbb{R}. \tag{2}$$

On peut alors estimer la qualité des simulations en mesurant une norme ou une distance entre $\mathcal{T}_V$ et S_V. Le travail du downscaling est de minimiser ces distances.

1.1 Introduction à la problématique du downscaling

Pour cette partie nous devons introduire les définitions de **fonction de répartition** et de **fonction de répartition empirique**.

Définition 3. *Soit X une variable aléatoire réelle, on appelle **fonction de répartition de** X, $\mathcal{F}_X : \mathbb{R} \to [0,1]$ la fonction vérifiant*

$$\mathcal{F}_X(x) = P(X \leqslant x). \tag{3}$$

Définition 4. *Soit Xune variable aléatoire réelle, on note f_X la **fonction de densité de** X,$f_X : \mathbb{R} \to [0,\infty[$ la fonction vérifiant*

$$f_X(x) = \lim_{t \to 0} \frac{P(x \leqslant X \leqslant x+t)}{t}. \tag{4}$$

Remarquons qu'un variables aléatoires ne possède pas nécessairement de fonction de répartition (notamment les variables aléatoires à valeurs discrètes). On peut cependant les étudier dans la théorie des distributions.

Définition 5. *Soient $X_1, X_2, ..., X_n$, n réalisations indépendantes d'une variable aléatoire réelle X, on appelle **fonction de répartition empirique de** X, la fonction $\mathcal{F}_n : \mathbb{R} \to [0,1]$ définie par*

$$\mathcal{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[X_i, +\infty)}(x). \tag{5}$$

Commençons par établir notre problématique dans le cas le plus simple où l'on cherche à prédire une variable V dans l'avenir alors que nous connaissons ses réalisations dans le passé à un endroit donné x. On peut alors considérer deux processus aléatoires à valeurs dans $\mathbb{R}$, $(X_t)_{t\in\mathbb{N}} = (\mathcal{S}_V(t,x))_{t\in\mathbb{N}}$ et $(Y_t)_{t\in\mathbb{N}} = (\mathcal{T}_V(t,x))_{t\in\mathbb{N}}$.

La problématique à laquelle nous cherchons de répondre est la suivante : connaissant $X_1, X_2, ..., X_n$ et $Y_1, Y_2, ..., Y_n$ les réalisations jusqu'au temps n ainsi que $X_{i_1}, X_{i_2}, ..., X_{i_m}$, on cherche une fonction $G : \mathbb{R} \to \mathbb{R}$ telle que les tirages $G(X_{i_1}), ..., G(X_{i_m})$ et $Y_{i_1}, ..., Y_{i_m}$ soient proches du point de vu de leur loi (nous éclaircirons ce point par la suite voir 2).

Autrement dit, en appelant X et Y les réalisations de $(X_t)_{t\in\mathbb{N}}$ et $(Y_t)_{t\in\mathbb{N}}$ sur $\{1, ..., n\}$ et X' et Y' les réalisations de $(X_t)_{t\in\mathbb{N}}$ et de $(Y_t)_{t\in\mathbb{N}}$ sur $\{i_1, ..., i_m\}$, on cherche à définir G à partir de X, Y et X' tel que $G_{X,Y,X'}$ minimise

$$d(\mathcal{F}_{G_{X,Y,X'}(X')}, \mathcal{F}_{Y'}),$$

où d est une distance définie sur les fonctions. Nous voulons aussi que $G_{X,Y,X'}$ respecte certaines propriétés. Par exemple, on veut que

Définition 6. *Soient X et Y deux variables aléatoires réelles et $G_{X,Y,X'} : \mathbb{R} \to \mathbb{R}$ une transformation, on dit que $G_{X,Y,X'}$ est **consistante vis à vis de X et de Y** si elle vérifie*

$$\mathcal{F}_{G_{X,Y,X}(X)} = \mathcal{F}_Y. \tag{6}$$

Dans la suite les transformations que nous considérerons seront toujours consistantes.

1.2 Cumulative Distribution Function transform (CDFt)

Nous allons ici présenter l'algorithme principal étudié et utilisé. On commencera par présenter l'algorithme de quantile-quantile permettant de comprendre l'esprit des transformations G affectées aux processus aléatoires.

1.2.1 Quantile-Quantile

Le quantile-quantile consiste simplement à définir $G_{X,Y}$ la transformation permettant de passer de l'une à l'autre.

Proposition 1. *Soit X et Y deux variables aléatoires réelles ayant des fonctions de répartition $\mathcal{F}_X$ et $\mathcal{F}_Y$ continues, alors $\mathcal{F}_Y^{-1}(\mathcal{F}_X(X))$ et Y suivent la même loi.*

Démonstration. Montrons que $\mathcal{F}_Y^{-1} \circ \mathcal{F}_X(X)$ et Y possède la même fonction de densité.

$$\mathcal{F}_{\mathcal{F}_Y^{-1} \circ \mathcal{F}_X(X)}(y) = \mathbb{P}(\mathcal{F}_Y^{-1}(\mathcal{F}_X(X)) \leqslant y) = \mathbb{P}(\mathcal{F}_X(X) \leqslant \mathcal{F}_Y(y)) = \mathcal{F}_Y(y). \tag{7}$$

Car $\mathcal{F}_X(X)$ suit une loi uniforme sur $[0, 1]$ si F_X est continue.

□

Le principe de l'algorithme **quantile-quantile** est de calculer la transformation $G = \mathcal{F}_Y^{-1} \circ \mathcal{F}_X$. Ainsi, $\mathcal{F}_{G(X)} = \mathcal{F}_Y$ et l'on définit alors $G_{X,Y,X'} = G$. L'une des limites de cette méthode est que le support de $f_{G(X)}$ est inclus dans celui de f_Y, alors les valeurs prises par $G(X')$ seront incluses dans $supp(f_Y)$. Nous aimerions que X ai aussi une influence sur ce résultat.

1.2.2 CDFt

L'algorithme de **Cumulative Distribution Function transfer** (CDFt) vise à remédier au problème des bornes en appliquant des transformations sur les lois X et Y.

CDFt avec régression linéaire :

Supposons que l'on ait des fonctions estimant la variance et de la moyenne de Y' à partir de X', s'exprimant sous la forme $\overline{f(X')} = \overline{Y'}$ ainsi que $\overline{g(X')} = \overline{\sigma(Y')}$ (on suppose ici X' et Y' comme des suites de variables aléatoires) [1]. On pose alors une condition supplémentaire sur $G_{X,Y,X'}$, imposant la conservation de la moyenne et de la variance. Formellement, on voudrait définir $G_{X,Y,X'}$ telle que quelque soient $\{i_1, ..., i_m\}$ un ensemble d'entiers consécutifs et X' et Y' les vecteurs des variables aléatoires des X_{i_j} et Y_{i_j} sur cet ensembles, on ait :

$$\overline{G_{X,Y,X'}(X')} = \overline{Y'}, \tag{8}$$

ainsi que

$$\sigma(G_{X,Y,X'}(X')) = \sigma(Y'), \tag{9}$$

et

$$G_{X,Y,X}(X) = Y, \quad \text{en loi.} \tag{10}$$

Concrêtement, nous allons faire des transformations sur les variables aléatoires pour avoir ces conditions là. On peut alors considérer la transformation

$$G_{X,Y,X'}(X') = \overline{g(X')} \frac{\mathcal{F}_Y^{-1} \circ \mathcal{F}_X(X') - \overline{\mathcal{F}_Y^{-1} \circ \mathcal{F}_X(X')}}{\sigma(\mathcal{F}_Y^{-1} \circ \mathcal{F}_X(X'))} + \overline{f(X')}.$$

1. On définit pas $\overline{X}$ la moyenne X et par $\sigma(X)$ son écart type

On peut vérifier facilement que $G_{X,Y,X'}$ ainsi défini respecte bien les trois propriétés précèdents. L'idée est alors de trouver les fonctions f et g par des méthodes de regression linéaires. Remarquons aussi que lorsque la loi est bornée cela rajoute encore une condition à $G_{X,Y,X'}$ et l'on ne peut avoir les conditions sur la moyenne et la variance, il faut alors choisir parmis l'une de ces deux conditions, c'est ce que nous faisons avec des précipitations par exemple.

1.3 Transport optimal

Remarquons que nous voulons à la fois prédire la précipitation et l'évapotranspiration, on peut alors considérer la variable aléatoire dans $\mathbb{R}^2$. On peut utiliser la même idée que celle décrite précédemment pour trouver une méthode permettant de corriger les biais statistiques introduite par les modèles de prédictions. Cette méthode a d'autant plus d'intérêt que la variable utile dans les modèles hydrologique est

$$\textit{pluie entrant dans le sol} = \textit{précipitation} - \textit{évapotranspiration},$$

Comme l'objectif final est de prédire des résultats hydrologiques sur le bassin du Little Washita, il semble particulièrement pertinent de considérer la loi conjointe précipitation, évapotranspiration.

La problématique du transport optimal a premièrement été introduite en par Gaspard Monge en 1781 puis a été développée par Kantorovitch en 1971 et ses travaux pour l'allocution des ressources lui ont valu un prix nobel d'économie en 1975.

1.3.1 Problématique

Nous donnons ici la formulation établie par Kantorovitch dans les années 70 qui a l'avantage d'inclure celle de Monge. Un lecteur intéressé par ce sujet pourra trouver un très bonne introduction dans le livre Villani (2003).

Considérons deux fonctions de répartitions pour des variables aléatoires U et V à valeurs dans X et Y, on appelle ces fonctions $\mathcal{F}$ et $\mathcal{G}$ et on appelle f et g leurs fonctions de densités. On cherche une mesure π sur $X \times Y$ satisfaisant

$$\int_Y d\pi(x,y) = f(x), \quad \int_X d\pi(x,y) = g(y),$$

et l'on cherche la mesure π satisfaisant l'équation précédente et minimisant la quantité

$$\mathcal{I}[\pi] = \int_{X\times Y} d(x,y)\, d\pi(x,y),$$

où d est une certaine distance définissant le coût de transport de x à y. Dans notre cas X et Y sont des variables aléatoires à valeurs dans $\mathbb{R}^2$. Nous voyons que le choix de la distance a une influence majeure sur les résultats obtenus. Alors on peu voir $d\pi(x,y)$ comme la quantité déplacée de x à y.

On peut alors calculer la fonction $\pi_{m,n}$ dans le cas fini, représentant la fonction empirique. On a $X_1, ..., X_n$ ainsi que $Y_1, ..., Y_m$ on cherche la matrice de transition $\pi_{i,j}$ $1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n$ telle que

$$\sum_{j=1}^{n} \pi_{i,j} = P(X = X_i) \text{ et } \sum_{i=1}^{m} \pi_{i,j} = P(Y = Y_j),$$

avec π minimisant

$$\mathcal{I}[\pi] = \sum_{i,j} d(X_i, Y_j)\pi_{i,j}.$$

La méthode d'obtention de la matrice π est donnée dans l'article Robin et al. (2019).

2 Analyse des résultats obtenus par downscaling

Concrètement nous allons faire de la validation croisée, nous allons apprendre sur 50% de nos données et faire nos prédictions. La partie à laquelle nous allons nous intéresser ici est la distance que nous utilisons pour évaluer nos prédictions. Contrairement à la manière habituelle de faire, consistant à estimer une distance entre chaque point prédit (souvent RMSE), en climatologie nous cherchons à comprendre la tendance générale. En effet, le paradigme d'évaluation en prévisions climatiques sur plusieurs années n'a pas l'ambition de prédire ponctuellement chaque prévision, mais il a pour objectif de décrire une tendance générale. On s'intéresse alors à des informations plus générales, c'est à dire que l'on travaille sur les lois de répartitions. Il faut alors réfléchir à des normes ou des distance pour évaluer la qualité de nos prédictions.

2.1 Tests basés sur les fonctions de répartition empiriques

Dans notre cas nous faisons des tests non-paramétriques, c'est à dire que l'on ignore tout des lois que nous comparons. Différentes méthodes pour tester l'égalité de lois sont connues, nous n'en développerons que deux. Le mémoire Éthier (2011) donne une présentations de principaux tests statistiques permettant d'évaluer si oui ou non à partir des réalisations $X_1, ..., X_n$, $Y_1, ..., Y_n$ de deux lois inconnues sont les mêmes.

Nous posons habituellement en statistiques deux hypothèses :

$$\mathcal{H}_0 : \mathcal{F}_X = \mathcal{F}_Y \quad et \quad \mathcal{H}_1 : \mathcal{F}_X \neq \mathcal{F}_Y,$$

où les égalités sur les lois sont en norme L^p. On suppose $\mathcal{H}_0$ et on définit une statistique sur $\|\mathcal{F}_X - \mathcal{F}_Y\|_{L^p(\mathbb{R})}$ permettant à partir de nos observations d'accepter ou de rejeter l'hypothèse $\mathcal{H}_0$. Les tests de Kolmogorov-Smirnov et Cramer-von Mises utilisent à-peu-près cette idée.

2.1.1 Quelques outils mathématiques

Proposition 2. *Soient $X_1, ..., X_n$ n réalisations d'une variable aléatoire réelle X et $\mathcal{F}_n$ sa fonction de répartition empirique nous avons*

$$E[\mathcal{F}_n] = F_X,$$

alors la fonction de répartition empirique est un estimateur sans biais de la lois de F.

Démonstration. C'est en effet évident puisque $\mathbb{1}_{[X,+\infty)}(x)$ suit une lois de Bernoulli de paramètre $\mathcal{F}(x)$ alors

$$E\Big[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{[X_i,+\infty)}(x)\Big] = \frac{1}{n}\sum_{i=1}^{n}E[\mathbb{1}_{[X_i,+\infty)}] = \mathcal{F}_X(x).$$

□

Théorème 2.1. *(Glivenko-Cantelli) Soient $\mathcal{F}_X$ et $\mathcal{F}_n$ respectivement la fonction de répartition et la fonction de répartition empirique. Alors*

$$\|\mathcal{F}_X - \mathcal{F}_n\|_\infty \xrightarrow[n\to\infty]{prob} 0 \tag{11}$$

Démonstration. (Cas où $\mathcal{F}_X$ est continue) On commence par remarquer que quelque soit x dans $\mathbb{R}$,

$$F_n(x) \xrightarrow[n\to\infty]{p.s.} \mathcal{F}_X(x)$$

d'après la loi forte des grands nombres et la proposition (2). Pour q dans $\mathbb{Q}$, on définit

$$\Omega_q = \{\omega \in \Omega | \lim_{n\to\infty} \mathcal{F}_n(q) = \mathcal{F}_X(q)\},$$

d'après ce que nous avons dit, sa mesure pour la probabilité $P(\Omega_q) = 1$ comme $\mathbb{Q}$ est dénombrable nous avons

$$P\Big(\bigcap_{q\in\mathbb{Q}}\Omega_q\Big) = 1.$$

Alors, comme $\mathbb{Q}$ est dense dans $\mathbb{R}$ et que $\mathcal{F}_X$ et les $(\mathcal{F}_n)_{n\in\mathbb{N}}$ sont continues on peut assurer que

$$\|\mathcal{F}_X - \mathcal{F}_n\|_\infty \xrightarrow[n\to\infty]{prob} 0.$$

□

Le cas où $\mathcal{F}_X$ n'est pas continue est géré par Durrett (2019) (ex 7.2 chap 1).

2.1.2 Kolmogorov-Smirnov

Le test d'ajustement de Kolmogorov-Smirnov est l'un des plus utilisé pour tester l'égalité de deux lois de probabilités. Dans le contexte de l'égalité de lois de probabilité, la statistique de test est

$$K_{n,m} = \sqrt{\frac{nm}{n+m}} \|\mathcal{F} - \mathcal{F}_n\|_\infty.$$

La suite de variables aléatoires $K_{n,m}$ converge vers une variable aléatoire K dont la fonction de survie est donnée par :

$$Q(x) = P(K > x) = \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2(jx)^2) \tag{12}$$

On peut alors l'approximer avec les premiers termes de la série pour construire notre test statistique. La démonstration de ce théorème peut être trouvée dans le livre Fisz (1963)(chap 12.5).

2.1.3 Cramér-von Mises

On considère ici deux fonctions de répartitions $\mathcal{F}$ et $\mathcal{G}$ continues. Nous voulons tester les hypothèses

$$\mathcal{H}_0 : \mathcal{F} = \mathcal{G} \quad \mathcal{H}_1 : \mathcal{F} \neq \mathcal{G},$$

Dans les tests proposés les statistiques sont construites à partir de deux échantillons indépendants (ce qui n'est pas notre cas). On définit aussi la fonction de répartition empirique $\mathcal{F}_n$.

Nous avons donc n réalisations de X et m réalisations de Y de lois de répartitions $\mathcal{F}$ et $\mathcal{G}$. La statistique du test est définie par

$$C_{n,m} = \frac{nm}{n+m} \int_{\mathbb{R}} \big[\mathcal{F}_n(x) - \mathcal{G}_m(x)\big]^2 \mathrm{d}\mathcal{H}_{m,n}(x), \tag{13}$$

avec,

$$\mathcal{H}_{n,m} = \frac{n}{n+m}\mathcal{F}_n + \frac{m}{n+m}\mathcal{G}_m. \tag{14}$$

$\mathcal{H}_{n,m}$ est alors la fonctions de répartitions empirique d'une variable aléatoire Z construite à partir des $n+m$ réalisations indépendantes $X_1, ..., X_n, Y_1, ..., Y_m$. On peut simplement réécrire la valeur $C_{n,m}$

$$C_{n,m} = \frac{mn}{(m+n)^2} \sum_{i=1}^{m+n} \big(\mathcal{F}_n(Z_i) - \mathcal{G}_m(Z_i)\big)^2 \tag{15}$$

Lemme 1. *On peut simplifier cette formule en supposant que les $(X_i)_{i\in\{1,...,n\}}$ et $(Y_i)_{i\in\{1,...,m\}}$ sont triés on a :*

$$C_{n,m} = \frac{1}{nm(m+n)}\Big[n\sum_{i=1}^{n}(R_{X_i} - i)^2 + m\sum_{i=1}^{m}(R_{Y_i} - i)^2\Big] - \frac{4nm-1}{6(m+n)}. \tag{16}$$

où R_{X_i} est le rang de X_i dans $X_1, ..., X_n, Y_1, ..., Y_n$ autrement dit

$$R_{Z_i} = Card(\{z \in Z, z \leqslant Z_i\}).$$

Note 1. *La démonstration de cette formule se trouve dans l'indexe 1 et la formulation de cette égalité diffère de celle contenue dans la thèse de Éthier (2011)(sec 2.3.2) qui contient une erreur.*

Cette formulation a le bon goût de nous indiquer que la statistique ne dépend pas de la loi. On peut calculer simplement sa statistique sous l'hypothèse $\mathcal{H}_0$ pour la loi uniforme sur $[0, 1]$ et ainsi retrouver les quantiles dans l'article de Büning (2002). Nous voyons que l'idée du test est aussi de pondérer la différence des fonctions de répartition empiriques par des (l'intégration selon $\mathcal{H}$). Ainsi, si le test de Kolmogorov-Smirnov est sensible aux outliers, le second l'est beaucoup moins lorsque les échantillons sont suffisamment grands.

3 Prédictions climatiques

Un grosse partie de notre travail aura été de prédire les variables d'évapotranspiration et de précipitation, nous verrons dans la section **??** que nous aurons seulement besoin de ces variables pour simuler le bassin hydrologique.

3.1 Les données NARR et la méthodologie

Les données NARR (North American Regional Reanalysis) couvrent l'entièreté du continent nord américain. La méthode de projection pour passer de $\mathcal{S}(\mathbb{R}^3)$ à $\mathbb{R}^2$ et ce qu'on appelle la projection Lambert conformale.

3.1.1 Projection Lambert conformale

4 Upscaling

L'upscaling est un co

5 Indexe 1

5.1 Lemme 1 sur la statistique de Kantorovitch

Soient $(X_i)_{i\in\{1,...,n\}}$ et $(Y_i)_{i\in\{1,...,m\}}$ des réalisations indépendante issues de variables aléatoires réelles X et Y. On appelle $\mathcal{F}_n$ et $\mathcal{G}_m$ les fonctions de répartitions empiriques définie à partir de ces réalisation et $\mathcal{H}_{m,n}$ la fonction de répartition empirique définie à partir de l'ensemble de ces réalisations $(Z_i)_{i\in\{1,...,m+n\}} = X_1, ..., X_n, Y_1, ..., Y_m$. Par la suite on considéra que tous les éléments sont triés dans leur ensemble $(i \leqslant j \Rightarrow E_i \leqslant E_j)$. Nous avons alors l'égalité suivante

$$C_{n,m} = \frac{nm}{n+m}\int_{\mathbb{R}}\big[\mathcal{F}_n(x)-\mathcal{G}_m(x)\big]^2\mathrm{d}\mathcal{H}_{m,n}(x) = \frac{1}{nm(m+n)}\Big[n\sum_{i=1}^{n}(R_{X_i}-i)^2+m\sum_{i=1}^{m}(R_{Y_i}-i)^2\Big]-\frac{4nm-1}{6(m+n)}. \quad (17)$$

où R_{X_i} est le rang de X_i dans $X_1, ..., X_n, Y_1, ..., Y_n$ autrement dit

$$R_{X_i} = Card(\{j, Z_j \leqslant X_i\}).$$

Notons que cette égalité transforme un problème d'analyse en un problème de dénombrement beaucoup plus simple. On rappelle la définition de l'intégrale par rapport à une fonction.

Définition 7. *Soient f une fonction continue par morceaux de $\mathbb{R}$ dans $\mathbb{R}$, soit g une fonction continue par morceaux on définit l'intégrale en appelant $x_{i,n} = i/n$*

$$\int_{\mathbb{R}} f(x)\,dg(x) = lim_{n\to\infty}\sum_{i\in\mathbb{Z}} f(x_{i,n})\big(g\big(x_{i,n}\big) - g\big(x_{i,n-1}\big)\big).$$

Démonstration. Commencons par montrer l'égalité

$$\frac{nm}{n+m}\int_{\mathbb{R}}\big[\mathcal{F}_n(x)-\mathcal{G}_m(x)\big]^2\mathrm{d}\mathcal{H}_{m,n}(x) = \frac{mn}{(m+n)^2}\sum_{i=1}^{m+n}\big(\mathcal{F}_n(Z_i)-\mathcal{G}_m(Z_i)\big)^2.$$

Ceci est vrai car en posant $\delta = \inf\{|Z_i - Z_j|, Z_i \neq Z_j\}$ alors quel que soit $n > n_0$ tel que $1/n_0 < \delta/2$ on a :

$$\sum_{i\in\mathbb{Z}}\big(\mathcal{F}_n(\frac{i}{n})-\mathcal{G}_m(\frac{i}{n})\big)^2\Big(\mathcal{H}_{m,n}\big(\frac{i}{n}\big)-\mathcal{H}_{m,n}\big(\frac{i-1}{n}\big)\Big) = \sum_{i=1}^{m+n}\big(\mathcal{F}_n(Z_i)-\mathcal{G}_m(Z_i)\big)^2,$$

on obtient donc directement l'égalité voulue.

Observons maintenant que $\mathcal{F}_n(X_i) = i/n$ et $\mathcal{G}_m(X_i) = (R_{X_i} - i)/m$ ainsi que $\mathcal{F}_n(Y_i) = (R_{Y_i} - i)/n$ et $\mathcal{G}_m(X_i) = i/m$. On peut alors réécrire $C_{n,m}$ en séparant la somme sur les X_i et Y_i

$$C_{n,m} = \frac{mn}{(m+n)^2}\Big[\sum_{i=1}^{n}\Big(\frac{i}{n}-\frac{R_{X_i}-i}{m}\Big)^2+\sum_{i=1}^{m}\Big(\frac{R_{Y_i}-i}{n}-\frac{i}{m}\Big)^2\Big]$$

$$= \frac{mn}{(m+n)^2}\Big[\frac{1}{m^2}\sum_{i=1}^{n}\Big(R_{X_i}-i\frac{m+n}{n}\Big)^2+\frac{1}{n^2}\sum_{i=1}^{m}\Big(R_{Y_i}-i\frac{m+n}{m}\Big)^2\Big]$$

Remarquons que $C_{n,m}$ est de la forme

$$C_{n,m} = \frac{mn}{(m+n)^2}\Big[\frac{C_1}{m^2} + \frac{C_2}{n^2}\Big],$$

Remarquons que C_1 et C_2 sont symétriques en n et m. On définit $\Sigma_1 = \sum_{i=1}^{n} R_{X_i}^2$, $\Sigma_2 = \sum_{i=1}^{m} R_{Y_i}^2$ et $\mathcal{S}_k = \sum_{i=1}^{k} i^2$ nous allons travailler sur l'expression

$$C_1 = \sum_{i=1}^{n} \Big(R_{X_i} - i\frac{m+n}{n}\Big)^2.$$

On la développe puis factorise pour obtenir

$$C_1 = \frac{m+n}{n}\sum_{i=1}^{n}(R_{X_i} - i)^2 - \frac{m}{n}\Sigma_1 + \frac{m(m+n)}{n^2}\mathcal{S}_n.$$

On obtient de la même manière

$$C_2 = \frac{m+n}{m}\sum_{i=1}^{m}(R_{X_i} - i)^2 - \frac{n}{m}\Sigma_2 + \frac{n(m+n)}{m^2}\mathcal{S}_m.$$

D'après ce qu'on a dit précédemment on a donc :

$$C_{n,m} = \frac{1}{nm(m+n)}\Big[n\sum_{i=1}^{n}(R_{X_i} - i)^2 + m\sum_{i=1}^{m}(R_{Y_i} - i)^2\Big] - \frac{\Sigma_1 + \Sigma_2}{(m+n)^2} + \frac{\mathcal{S}_n}{n(n+m)} + \frac{\mathcal{S}_m}{m(m+n)}.$$

On remarque $\Sigma_1 + \Sigma_2 = \mathcal{S}_{m+n}$ et que l'on a la première moitié de notre somme. Il ne reste plus qu'à développer l'expression

$$-\frac{\mathcal{S}_{m+n}}{(m+n)^2} + \frac{\mathcal{S}_n}{n(n+m)} + \frac{\mathcal{S}_m}{m(m+n)}$$

$$= -\frac{(m+n+1)(2m+2n+1)}{6(m+n)} + \frac{(n+1)(2n+1)}{6(m+n)} + \frac{(m+1)(2m+1)}{6(m+n)} = -\frac{4mn-1}{6(m+n)}$$

En regroupant nos deux résultats nous avons finalement :

$$C_{n,m} = \frac{1}{nm(m+n)}\Big[n\sum_{i=1}^{n}(R_{X_i} - i)^2 + m\sum_{i=1}^{m}(R_{Y_i} - i)^2\Big] - \frac{4nm-1}{6(m+n)}$$

□

Références

Büning, H. (2002). Robustness and power of modified lepage, kolmogorov-smirnov and crame´ r-von mises two-sample tests. *Journal of Applied Statistics*, 29(6) :907–924.

Durrett, R. (2019). *Probability : theory and examples*, volume 49. Cambridge university press.

Éthier, F. (2011). *À propos de divers tests statistiques pour l'égalité des lois*. PhD thesis, Université du Québec à Trois-Rivières.

Fisz, M. (1963). Probability theory and mathematical statistics.

Robin, Y., Vrac, M., Naveau, P., and Yiou, P. (2019). Multivariate stochastic bias corrections with optimal transport. *Hydrology and Earth System Sciences*, 23(2) :773–786.

Villani, C. (2003). *Topics in optimal transportation*. Number 58. American Mathematical Soc.