Contribution à l'évaluation d'un patrimoine iconographique
en vue d'applications aux pays d'Afrique subsaharienne
Mesure des disparités liées à la corrélation des caractères temporel et régional

par Stéphane Richemond*

Nous avons souvent été confrontés aux disparités des patrimoines iconographiques. Ce fut le cas pour les expositions que I&M avait préparées sur les villes de Bamako, Lomé, Ouagadougou… ou sur le Vietnam. Nous avons pensé qu'il n'était pas suffisant de faire le constat de l'absence d'images sur tel thème, tel quartier ou telle région, telle période, mais qu'il serait souhaitable de quantifier ces carences de façon objective à l'aide d'indices qui donneraient une mesure des disparités permettant de caractériser un patrimoine et de comparer des patrimoines différents. Cette idée, qui semblait ne pas avoir été explorée, nous a conduit à établir les formules présentées dans cet article et dans les deux précédentes livraisons de ce *Bulletin*[1]. Nous avions vu comment mesurer les disparités temporelles et régionales du patrimoine iconographique d'un pays[2]. Après nous être intéressés à la distribution d'une population d'images selon les deux caractères, la période *P* et le lieu (ou la région) *R*, ceux-ci étant pris isolément, nous présentons ici une mesure des disparités dues à leur corrélation.

I - Rappel sur la mesure des disparités temporelles d'une population d'images

Une population d'images couvre toujours de façon inégale les périodes considérées. Supposons que l'intervalle de temps étudié soit la réunion de m périodes successives. Soit N le nombre total d'images toutes périodes confondues. Notons n_i le nombre d'images appartenant à la période P_i et f_i la proportion n_i/N d'images appartenant à cette période. Selon le caractère P, la distribution (dite marginale) d'images peut être décrite dans le tableau suivant :

P_i	P_1	P_2		P_i		P_m	Total
f_i	f_1	f_2		f_i	.	f_m	1

Supposons les périodes P_i d'égale longueur, alors on peut considérer que la population d'images serait harmonieusement distribuée si chaque période contenait le même nombre d'images N/m soit la même proportion p_i=1/m d'images. Une telle distribution idéale peut être représentée dans le tableau suivant :

P_i	P_1	P_2		P_i			P_m	Total
p_i	1/m	1/m		1/m			1/m	1

Nous pouvons mesurer l'écart entre la distribution d'images et la distribution idéale de référence par l'indicateur suivant que nous avons appelé **indice de disparité temporelle**[3] :

***** IRHiS (Université de Lille) - srichemond@hotmail.com

[1] - Stéphane Richemond, "Contribution à l'évaluation d'un patrimoine iconographique en vue d'applications aux pays d'Afrique subsaharienne - Mesure des disparités temporelles", *Bulletin n°53*, Images & Mémoires, Été 2017.

- Stéphane Richemond, "Contribution à l'évaluation d'un patrimoine iconographique en vue d'applications aux pays d'Afrique subsaharienne - Mesure des disparités régionales", *Bulletin n°54*, Images & Mémoires, Automne 2017.

[2] Nous avions défini un tel patrimoine par l'ensemble des images (dessins, peintures, photographies,...) identifiables comme appartenant au pays considéré, où qu'elles soient localisées, même si elles ne sont pas à la disposition du public.

[3] Il s'agit de la distance euclidienne définie par le théorême de Pythagore généralisé. En divisant E_p par sa valeur maximum prise lorsque toutes les images sont concentrées sur une seule période, on obtient un indice compris entre 0 et 1.
Nous avions vu précédemment que nous pouvions aussi utiliser la distance du khi[2] définie par :

$$\sqrt{\sum_{i=1}^{m}\left(\frac{f_i - p_i}{\sqrt{p_i}}\right)^2}$$

$$E_P = \sqrt{\sum_{i=1}^{m}(f_i - \frac{1}{m})^2} \text{ où on a } \sum_{i=1}^{m} n_i = N \text{ et } \sum_{i=1}^{m} f_i = 1$$

II - Rappel sur la mesure des disparités régionales d'une population d'images

Il nous a semblé naturel de considérer une distribution uniforme comme référence pour la mesure des disparités temporelles ce qui n'est pas justifié pour la mesure des disparités régionales car les régions ont des superficies et des populations différentes. Nous avons pensé pertinent de considérer des proportions de référence proportionnelles aux superficies des régions, ou encore à leurs populations. Supposons que l'étendue E du pays soit la réunion de p régions R_j. Notons n_j le nombre d'images concernant la région R_j et $f_j = n_j/N$ la proportion correspondante. La distribution (dite marginale) d'images est alors la suivante :

R_j	R_1	R_2		R_j		R_p	Total
f_j	f_1	f_2		f_j	.	f_p	1

Notons S la superficie de l'ensemble des régions considérées, et S_j la superficie de la région R_j, de même A la population totale et A_j celle de la région R_j. Prenons pour proportion de référence $p_j = S_j/S$. Nous pouvons mesurer l'écart entre la distribution d'images et la distribution de référence par l'expression suivante que nous avons appelée **indice de disparité régionale relatif aux superficies** :

$$E_{R/S} = \sqrt{\sum_{j=1}^{p}(f_j - \frac{S_j}{S})^2}$$

Définissons maintenant la proportion de référence par $p_j = \frac{A_j}{A}$. Nous avons mesuré l'écart entre la distribution d'images et la distribution de référence par l'expression suivante que nous avons appelé **indice de disparité régionale relatif aux populations** :

$$E_{R/A} = \sqrt{\sum_{j=1}^{p}(f_j - \frac{A_j}{A})^2}$$

Par ailleurs, comme annoncé précédemment, on adoptera plus volontiers la formule de la distance du khi^2 pour chacun de ces indices. Pour prendre en compte à la fois les superficies et les proportions pour chaque région, nous pourrons prendre des proportions de référence du type :

$$p_j = \alpha\frac{S_j}{S} + (1-\alpha)\frac{A_j}{A} \text{ où } \alpha \in [0,1]$$

III - Tableau de contingence

On s'intéresse maintenant à l'étude une population d'images sous l'angle simultané des deux caractères la période et la région. On notera n_{ij} (respectivement f_{ij}) le nombre (respectivement la proportion) d'images qui possèdent à la fois la modalité P_i du caractère *P* et la modalité R_j du caractère *R*. On a ainsi :

$$f_{i.} = \sum_{j=1}^{p} f_{ij} \text{ ; de même } f_{.j} = \sum_{i=1}^{m} f_{ij} \text{ et } \sum_{i=1}^{m}\sum_{j=1}^{p} f_{ij} = 1$$

où le point en indice signifie que l'on a effectué une sommation sur l'indice remplacé.
On peut ranger les effectifs ou les proportions dans un tableau à double entrée :

	R_1	R_2		R_j		R_p	$f_{i..}$
P_1	f_{11}	f_{12}		f_{1J}		f_{1p}	$f_{1.}$
P_2	f_{21}	f_{22}		f_{2i}		f_{2p}	$f_{2.}$
P_i	f_{i1}	f_{i2}		f_{ij}		f_{ip}	$f_{i.}$
P_m	$f_{m.}$	$f_{m.}$		f_{mj}		$f_{m.}$	$f_{m.}$
$f_{.j}$	$f_{.1}$	$f_{.2}$		$f_{.j}$		$f_{.p}$	1

Ainsi, on a porté à l'intersection de la ième ligne et de la jème colonne la proportion f_{ij} des images qui appartiennent à la fois à la période P_i et à la région R_j. Nous avions déjà insisté sur le grand intérêt qu'il y a à remplir ce tableau qui met en évidence les disparités des caractères en un coup d'œil.

IV - Distributions conditionnelles

On peut aussi étudier la population d'images sous l'angle d'un caractère, les modalités de l'autre étant fixée. Par exemple, on peut étudier le caractère *P* dans le cas où la modalité R_j de *R* est fixée. On obtient le tableau :

P_i	P_1	P_2		P_i		P_m	Total
f_i^j	f_1^j	f_2^j		f_i^j	.	f_m^j	$f_.^j$

où on a noté $f_i^j = n_{ij.}/n_{.j}$ ce qui se lit « f_i si j ».

On peut donc étudier, sous l'angle du caractère *P*, p disstributions à j fixé. De même, on peut étudier, sous l'angle du caractère *R*, m distributions à i fixé.

Il est clair que caractère *P* est indépendant du caractère *R* si la proportion d'individus possédant la modalité P_i de *P* reste constante pas lorsque j varie. En supposant que *P* et *R* soient les seuls caractères, on a alors :

$$f_i^j = \text{cte } \forall j \in \{1, 2, \ldots p\} \text{ soit } f_i^j = f_{i.}$$

On montre aisément que l'indépendance est une notion réciproque. En effet :

$$f_{ij} = \frac{n_{ij}}{N} = \frac{n_{ij}}{n_{.j}} \cdot \frac{n_{.j}}{N} = f_i^j f_{.j} = f_{i.} \cdot f_{.j} \text{ et } f_{ij} = f_j^i \cdot f_{i.}. \text{ On en déduit : } f_j^i = f_{.j}.$$

On observe alors que les lignes sont proportionnelles entre elles, de même que les colonnes le sont aussi.

V - Disparités des distributions conditionnelles

Nous avons déjà étudié les disparités des distributions marginales de *P* et *R*. Les indices E_R et E_P ne sont cependant pas suffisants pour caractériser toutes les disparités d'une population. En effet, imaginons le cas extrême où : $f_{i.}=f_{.j}\ \forall i \in \{1, 2, \ldots m\}, \forall j \in \{1, 2, \ldots p\}$, $f_{ij}=0$ si $i \neq j$, soit, par exemple, à la situation du tableau suivant :

	R_1	R_2	R_3	R_4	$n_{i.}$
P_1	0,25	0	0	0	0,25
P_2	0	0,25	0	0	0,25
P_3	0	0	0,25	0	0,25
P_4	0	0	0	0,25	0,25
$n_{.j}$	0,25	0,25	0,25	0,25	1

Alors que les distributions marginales de *P* et *R* montrent des disparités nulles ($E_R = E_P = 0$), nous avons une disparité maximum des distributions conditionnelles. En effet, toutes les images de la période P_1 sont concentrées sur la région R_1, et ainsi de suite. Ce cas correspond au maximum de corrélation (liaison fonctionnelle réciproque des deux caractères *P* et *R*.).

La disparité des distributions conditionnelles peut être mesurée par la distance entre le tableau de corrélation et le tableau correspondant à l'indépendance des caractères *P* et *R* qui est celui de la population de référence idéale. Le tableau correspondant à l'indépendance des caractères se construit aisément à partir de la connaissance des deux distributions marginales grâce à la relation :

$$f_{ij} = f_{i.}f_{.j} \ \forall i \in \{1, 2, \ldots m\}, \forall j \in \{1, 2, \ldots p\} \quad \text{qui s'écrit encore } n_{ij} = n_{i.}f_{.j}.$$

ce qui donnerait ici le tableau de référence (comprenant des proportions égales dans ce cas particulier) :

	R_1	R_2	R_3	R_4	$f_{i.}$
P_1	0,0625	0,0625	0,0625	0,0625	0,25
P_2	0,0625	0,0625	0,0625	0,0625	0,25
P_3	0,0625	0,0625	0,0625	0,0625	0,25
P_4	0,0625	0,0625	0,0625	0,0625	0,25
$f_{.j}$	0,25	0,25	0,25	0,25	1

D'une façon plus générale, nous définirons **l'indice de disparité des distributions conditionnelles** par :

$$E_{PR} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{p}\left(f_{ij} - f_{i.}f_{.j}\right)^2}$$

qui correspond à une mesure de la distance entre les deux tableaux précédents.

En pratique, les proportions du patrimoine iconographique ne sont pas connues et E_{PR} n'est donc pas calculable par la formule ci-dessus. Nous utiliserons cette formule en remplaçant les proportions par celles déterminées sur un échantillon aléatoire d'images, comme nous avions procédé lors des études précédentes concernant les disparités temporelles et régionales. Il serait souhaitable que l'échantillon soit le plus grand possible, mais il est surtout important que l'échantillon, qui est généralement une collection, n'ait pas été constitué avec des images appartenant à une période, une région ou un thème privilégié.

La quantité obtenue à partir des proportions f_{ij}^* déterminées sur l'échantillon aléatoire :

$$E_{PR}^* = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{p}\left(f_{ij}^* - f_{i.}^*f_{.j}^*\right)^2}$$

est une variable qui est un estimateur asymptotiquement sans biais[4] de E_{PR}, condition nécessaire mais non suffisante à sa convergence en probabilité vers E_{PR} lorsque la taille de l'échantillon n croît.

Conclusion

Cette étude ouvre la voie à des développements plus importants. En particulier, nous n'avons pas abordé la mesure des disparités thématiques pour lesquelles la définition d'une population de référence est délicate. Une réflexion sur la détermination d'un indice plus synthétique pourrait être menée. La question d'établir un échantillon aléatoire à partir de collections d'images n'est pas non plus un problème simple.

4

$$\text{Ecrivons } \sum_{i=1}^{m}\sum_{j=1}^{p}\left(f_{ij}^* - f_{i.}^*f_{.j}^*\right)^2 \text{ sous la forme : } \sum_{i=1}^{m}\sum_{j=1}^{p}\left((f_{ij}^* - f_{ij}) + \left(f_{ij} - f_{i.}f_{.j}\right) + (f_{ij}f_{ij} - f_{i.}^*f_{.j}^*)\right)^2$$

que nous développerons, obtenant ainsi la somme de six termes dont l'espérance mathématique tend vers E_{PR} lorsque n croît. Pour l'établir, on aura recours à :

$$E\left(f_{i.}^*f_{.j}^* - f_{i.}f_{.j}\right) = \mathrm{Cov}\left(f_{i.}^*, f_{.j}^*\right) \leq \sigma(f_{i.}^*)\sigma\left(f_{.j}^*\right) = \frac{1}{n}\sqrt{(1 - f_{i.})(1 - f_{.j})f_{i.}f_{.j}} \quad \text{qui tend vers 0 avec n.}$$

Ici Cov représente la covariance et σ représente l'écart-type.