Contribution à l'évaluation d'un patrimoine iconographique
en vue d'applications aux pays d'Afrique subsaharienne
Mesure des disparités régionales

par Stéphane Richemond*

Cet article est le second d'une série consacrée à l'évaluation d'un patrimoine iconographique. Il mobilise des données mathématiques, alors que notre Bulletin *est plutôt situé sur une approche liée aux* ***sciences humaines et sociales.*** *Néanmoins, ces sciences humaines utilisent régulièrement des outils et modèles mathématiques, statistiques, appartenant aux* ***sciences "exactes".*** *Il nous a donc paru intéressant et utile de publier cette contribution.*

J. M. Andrault

Nous nous intéressons ici à évaluer le patrimoine iconographique d'un pays défini, comme déjà proposé[1], par l'ensemble des images (dessins, peintures, photographies, cartes postales...) identifiables comme appartenant au pays considéré, où qu'elles soient localisées, même si elles ne sont pas à la disposition du public.

Deux qualités caractérisent principalement un tel patrimoine : la quantité d'images qu'il comprend et la distribution de celles-ci selon divers caractères. Bien que nous ne puissions améliorer ces caractéristiques, nous souhaiterions que la quantité d'images soit importante et que leur distribution soit harmonieuse.

Trois caractères semblent prioritaires pour contribuer à définir une image : la période *P*, le lieu (ou la région) *R*, le thème *T*. Selon le caractère, la population d'images couvre toujours de façon inégale le pays considéré. Une image peut, selon le caractère considéré, avoir plusieurs modalités. On peut décomposer l'intervalle de temps étudié en m sous-périodes P_i. On supposera que le pays est partagé en p régions distinctes R_j.

I – Généralités et notations

Nous avions déjà étudié dans la précédente livraison de ce *Bulletin* la mesure des disparités temporelles d'une population d'images. Nous proposons d'appliquer ici une démarche similaire à la mesure des disparités régionales. Celles-ci s'expliquent en partie par le fait que les premières photographies ont été prises par des Européens, et donc là où ils étaient présents tels les lieux où étaient implantés l'administration coloniale, les missions catholiques (Koupéla en Haute-Volta, Tahoua au Niger), ou les camps militaires (Kati au Soudan).

Les modalités des caractères *R* sont supposées incompatibles en ce sens qu'une image ne peut appartenir à la fois à deux régions[2]. Elles doivent être exhaustives ce qui signifie qu'à chaque image on peut associer une région, et bien sûr une seule du fait de leur incompatibilité.

On notera N la taille de la population d'images et n_j le nombre d'images possédant la modalité R_j de *R*. La proportion d'images appartenant à la région R_j est notée f_j. On a alors $f_j=n_j/N$. Nous noterons 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.

II - Descriptions marginales[3] d'une population d'images selon les régions – Mesure des disparités

On peut étudier la population d'images selon les divers caractères. Selon le caractère *R*, on obtient :

R_j	R_1	R_2		R_j		R_p	
n_j	n_1	n_2		n_j		n_p	N
f_j	f_1	f_2		f_j		f_p	1

***** srichemond@hotmail.com

[1] "Contribution à l'évaluation d'un patrimoine iconographique en vue d'applications aux pays d'Afrique subsaharienne - Mesure des disparités temporelles", *Bulletin n°53*, p. 37-40, Images & Mémoires, été 2017.

[2] Une image peut être plurithématique, auquel cas on pourra la considérer plusieurs fois, comme autant d'images que de thèmes. En revanche, elle ne peut appartenir qu'à une seule période et une seule région.

[3] Contrairement aux distributions conditionnelles de *R* étudiées sous la contrainte que des modalités des caractères *P* et *T* soient fixées, les distributions marginales sont celles qui concernent la population entière d'images.

Il n'est pas simple de définir une distribution idéale de référence par rapport à laquelle il conviendrait de mesurer la distance à la distribution étudiée. Il vient naturellement à l'idée de fixer des proportions d'images de la distribution idéale qui soient égales à celles des populations correspondant à chaque région, ou encore à l'aire de la région rapportée à la superficie du pays. On définit la proportion de référence $p_j = \frac{S_j}{S}$.

On obtient alors le tableau :

R_j	R_1	R_2		R_j		R_p	
$\frac{S_j}{S}$	$\frac{S_1}{S}$	$\frac{S_2}{S}$		$\frac{S_j}{S}$	.	$\frac{S_p}{S}$	1

Pour chaque région R_j, l'écart entre la proportion réelle et la proportion idéale d'images est : $e_j = f_j - \frac{S_j}{S}$.

Nous mesurerons l'écart entre la distribution d'images et la distribution idéale[4] par l'expression suivante que nous pouvons appeler indice de disparité régionale[5] :

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

Choisir des proportions proportionnelles aux superficies donne une importance exagérée à des régions désertiques en mettant sur un même plan les régions peu peuplées et très peuplées. Soit A_j la population d'habitants de la région R_j, à l'origine de la période étudiée par exemple. Définissons maintenant la proportion de référence par : $p_j = \frac{A_j}{A}$. On obtient le tableau :

R_j	R_1	R_2		R_j		R_p	
$\frac{A_j}{A}$	$\frac{A_1}{A}$	$\frac{A_2}{A}$		$\frac{A_j}{A}$	.	$\frac{A_p}{A}$	1

Pour chaque région R_j, l'écart entre la proportion réelle et la proportion idéale d'images[6] est : $e_j = f_j - \frac{A_j}{A}$.

Pour mesurer l'écart entre la distribution d'images et la distribution idéale nous proposerons l'indice de disparité régionale :

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

À titre d'illustration considérons un ensemble de trois régions R_1, R_2 et R_3, comprenant respectivement 50 000, 450 000 et 500 000 habitants. Supposons que nous disposions de 1 000 images dont 10 appartiennent à la région R_1, 290 à la région R_2 et 700 à la région R_3. Nous obtenons le tableau suivant :

---

[4] On a bien entendu : $\sum_{j=1}^{p} e_j = 0$

[5] Nous aurions pu aussi retenir la formule : $E_P = \sum \left| f_j - \frac{S_j}{S} \right|$ mais elle ne se prête pas bien aux calculs formels du fait de la présence de valeurs absolues.

[6] Comme en 4, la somme de ces écarts est nulle.

	R_1	R_2	R_3	Total
Population	50 000	450 000	500 000	1 000 000
A_j/A	0,05	0,45	0,50	1
Nombre d'images	10	290	700	1.000
Proportions observées f_j	0,01	0,29	0,7	1
f_j-A_j/A	-0,04	-0,16	0,20	0
$(f_j$-$A_j/A)^2$	0,0016	0,0256	0,04	0,0672

L'indice de disparité régional[7] $E_{R/A}$ est égal à 0,2592 (racine de 0,0672).

L'indice ainsi calculé est la distance euclidienne dans l'espace R^p entre les points (f_1, f_2, …, f_j,…, f_p) et (A_1/A, A_2/A, …, A_j/A,…, A_p/A).

Afin de prendre en compte à la fois la population et la superficie de chaque région, on peut adopter un indice de la forme :

$$E_{R/SA} = \sqrt{\sum_{j=1}^{p}(f_j - \alpha\frac{S_j}{S} - (1-\alpha)\frac{A_j}{A})^2} \text{ où } \alpha \in [0,1]$$

Par ailleurs, choisir des proportions égales à celles des proportions d'origine pose le problème que celles-ci peuvent avoir changé et ont été souvent modifiées par un fort exode rural. À titre d'exemple, la ville de Bamako, devenue capitale du Soudan français en mai 1908, a vu sa population passer en un siècle de quelques milliers à un million d'habitants. Elle a supplanté la ville de Kayes qui était alors la plus importante. Il en est de même de la ville d'Ouagadougou qui a supplanté celle de Bobo-Dioulasso dans les années 1950 avec l'arrivée du train, ou de la ville de Lomé qui a supplanté celle d'Anecho dans les années 1900 avec la construction du wharf. Il est donc clair que l'iconographie de Kayes est plus intéressante que celle de Bamako en 1900, alors que c'est

[7] On remarque que l'écart entre la proportion d'images et la proportion de référence dans la région R_1 (-0,04) est très faible en valeur absolue et apporte donc une faible contribution à la valeur de l'indice de disparité régionale ainsi calculé, alors qu'il est important en valeur relative : $\frac{0,01-0,05}{0,05} = -0,8$. Le calcul de l'indice de disparité appliqué aux valeurs relatives conduit à :

$$\sqrt{\sum_{j=1}^{p}(\frac{f_j - A_j/A}{A_j/A})^2} = 0,96$$

Cependant, dans ce résultat, l'écart provenant de la région R_1 apporte de loin la plus forte contribution à ce résultat. Un bon compromis semble être l'utilisation de la distance du Khi^2 définie par :

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

En reprenant l'exemple précédent, on obtient alors le tableau :

	R_1	R_2	R_3	Total
A_j/A	0,05	0,45	0,50	1
Proportions observées f_j	0,01	0,29	0,7	1
f_j-A_j/A	-0,04	-0,16	0,20	0
$(f_j$-$A_j/A)^2$	0,0016	0,0256	0,04	0,0672
$(f_j$-$A_j/A)^2/(A_j/A)$	0,032	0,05689	0,08	0,168

On trouve pour distance du Khi^2 $\sqrt{0,168}$ =0,41.

l'inverse dans les années 1950. C'est pourquoi il semble pertinent de ne pas adopter, pour la distribution de référence, des proportions égales uniquement à celles des populations de la période d'origine, mais de prendre en compte l'évolution des populations au cours de la période considérée.

Écrivons que l'intervalle de temps étudié est la réunion de deux sous-intervalles, le premier comprenant les périodes $\{P_1, P_2,.., P_s\}$,, le second comprenant les périodes $\{P_{s+1}, P_{s+2},.., P_m\}$. Notons n_{ij} (respectivement f_{ij}) le nombre (respectivement la proportion) d'images appartenant à la fois à la période P_i et à la région R_j.

Posons :

$$N_1 = \sum_{i=1}^{s}\sum_{j=1}^{p} n_{ij} \qquad N_2 = \sum_{i=s+1}^{m}\sum_{j=1}^{p} n_{ij} \qquad A^1 = \sum_{j=1}^{s} A_j^1 \qquad A^2 = \sum_{j=s+1}^{m} A_j^2$$

On a alors : $N= N_1+N_2$ et $Aj = A_j^1+A_j$

Les effectifs de la population de référence peuvent être consignés dans les tableaux suivants :

Tableau 1 (correspondant à $i \leq s$)

R_j	R_1	R_2		R_j		R_p	Total
$N_1\frac{A_j^1}{A^1}$	$N_1\frac{A_1^1}{A^1}$	$N_1\frac{A_2^1}{A^1}$		$N_1\frac{A_j^1}{A^1}$		$N_1\frac{A_p^1}{A^1}$	N_1

L'écart entre la population de référence décrite dans le tableau 1 et la population réelle d'images anciennes d'effectif N_1 peut être mesuré par :

$$E_{R/A}^1 = \sqrt{\sum_{j=1}^{p}\left(\left(\sum_{i=1}^{s} f_{ij}\right) - \frac{N_1}{N}\frac{A_j^1}{A^1}\right)^2}$$

Tableau 2 (correspondant à $i > s$)

R_j	R_1	R_2		R_j		R_p	Total
$N_2\frac{A_j^2}{A^2}$	$N_2\frac{A_1^2}{A^2}$	$N_2\frac{A_2^2}{A^2}$		$N_2\frac{A_j^2}{A^2}$		$N_2\frac{A_p^2}{A^2}$	N_2

L'écart entre la population de référence décrite dans le tableau 2 et la population réelle d'images récentes d'effectif N_2 peut être mesuré par :

$$E_{R/A}^2 = \sqrt{\sum_{j=1}^{p}\left(\left(\sum_{i=s+1}^{m} f_{ij}\right) - \frac{N_2}{N}\frac{A_j^2}{A^2}\right)^2}$$

Bien entendu, nous pourrions prendre un nombre plus important de sous-périodes mais la difficulté pour le calcul de ces indices est de déterminer les populations d'habitants de chaque région au début de chaque sous-période. Il est aussi intéressant d'avoir un seul indice global pour la population d'images plutôt qu'un indice propre à chaque sous-période. On obtient après calculs un indice de disparité régionale[8] :

8

En effet, développons $\sum_{j=1}^{p}\left(\sum_{i=1}^{n} f_{ij} - \frac{N_1}{N}\frac{A_j^1}{A^1} - \frac{N_2}{N}\frac{A_j^2}{A^2}\right)^2 = \sum_{j=1}^{p}\left(\left(\sum_{i=1}^{s} f_{ij} - \frac{N_1}{N}\frac{A_j^1}{A^1}\right) + \left(\sum_{i=s+1}^{p} f_{ij} - \frac{N_2}{N}\frac{A_j^2}{A^2}\right)\right)^2$

On obtient en développant le carré : $\sum_{j=1}^{p}\left(\sum_{i=1}^{s} f_{ij} - \frac{N_1}{N}\frac{A_j^1}{A^1}\right)^2 + \sum_{j=1}^{p}\left(\sum_{i=s+1}^{m} f_{ij} - \frac{N_2}{N}\frac{A_j^2}{A^2}\right)^2$

$$E_{R/A} = \sqrt{\sum_{j=1}^{p}\left((\sum_{i=1}^{s} f_{ij} - \frac{N_1}{N}\frac{A_1^1}{A^1}) ^2 + (\sum_{i=s+1}^{m} f_{ij} - \frac{N_2}{N}\frac{A_1^2}{A^2}) ^2 \right)}$$

Ceci n'est autre que la somme des carrés des deux coefficients $E_R{}^1$ et $E_R{}^2$ précédents. Soit :

$$E_{R/A} = \sqrt{(E_{R/A}^1)^2 + (E_{R/A}^2)^2}$$

Cette formule peut être étendue à un plus grand nombre d'intervalles de temps successifs.

III - Traitement des données en cas de collections d'images incomplètes

Le patrimoine iconographique d'un pays restera toujours en partie inconnu. Quand il est peu important il est possible de réaliser des collections d'images qui soient « presque exhaustive ». Si la collection d'images dont nous disposons est de quelques centaines d'unités, alors nous sommes éloignés de l'exhaustivité souhaitée et les calculs proposés ne sont que des estimations. Supposons par exemple que nous souhaitions avoir des informations sur le patrimoine iconographique d'un pays à partir d'une collection de n (=500) images, alors que l'exhaustivité en exigerait une dizaine de milliers. Une façon de procéder serait de considérer que notre collection est un échantillon aléatoire prélevé dans la population totale. Ceci ne sera possible que si la constitution de la collection n'a pas obéi à des critères particuliers autres que l'appartenance au pays considéré.

Dans ce cas, s'il convient d'avoir une estimation de l'indice :

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

où p_j est la proportion inconnue du patrimoine iconographique des images relatives à la région R_j, la première idée est de calculer l'indice de disparité sur l'échantillon :

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

où f_j est la proportion, calculée sur l'échantillon, des images appartenant à la région R_j.

f_j est une variable aléatoire[9] qui prendra sur chaque échantillon possible une valeur différente, inconnue

---

car la sommation sur j de chaque terme du double produit est nulle.

[9] Supposons que nous nous intéressions à l'estimation de la proportion d'images p_j de la région R_j de la population totale à partir de cet échantillon aléatoire de n unités. Il est clair que si nous prenons au hasard une image dans la population totale, celle-ci aura une probabilité p_j d'appartenir à la région R_j et une probabilité 1-p_j de ne pas lui appartenir. Répétons l'expérience n fois, avec remise, afin d'obtenir un échantillon. Notons f_j la proportion d'images de l'échantillon appartenant à la région R_j. Cette proportion est une variable aléatoire, car elle prendrait une valeur différente et inconnue d'avance pour chaque échantillon aléatoire de taille n. Cette proportion f_j, destinée à estimer la proportion inconnue p_j, suit une loi de probabilité binomiale B(n, p_j). Elle est un estimateur sans biais et convergent de p_j. Ses probabilités, son espérance mathématique et sa variance sont données par :

$$\Pr\left\{f_j = \frac{k}{n}\right\} = C_n^k p_j^k (1-p_j)^{n-k} \;;\; E(f_j) = p_j \text{ et } V(f_j) = \frac{p_j(1-p_j)}{n}$$

Nous avons supposé le tirage indépendant, alors qu'il est en général exhaustif ce qui impliquerait l'usage de la loi hypergéométrique, plus compliquée mais dont la variance est plus faible. Le choix fait est donc sans risque accru.

d'avance. Il est clair que l'estimateur E_R^* de E_R est aussi une variable aléatoire. On montre sans difficulté que l'espérance mathématique de E_R^* tend vers E_R lorsque n croît[10] et que la variance de E_R^* tend vers 0 avec n. La réalisation de ces deux conditions est suffisante pour que l'estimateur E_R^* converge en probabilité vers E_R. Nous retiendrons donc la valeur prise par E_R^* sur l'échantillon comme estimation de E_R.

IV - Amélioration des indices

On peut améliorer les indices de disparité temporelle ou régionale en divisant ceux-ci par la valeur qu'ils prendraient si la disparité était maximale, c'est-à-dire s'il y avait concentration des effectifs sur une seule période ou une seule région. Dans ce cas, l'indice obtenu, qui est égal à 0 en cas de disparité uniforme, sera égal à 1 en cas de disparité maximale. Il est bien sûr d'autant plus proche de 1 que la disparité est forte.

[10] Écrivons :

$$E_{R/A}^{*2} = \sum_{j=1}^{p}(f_j - \frac{A_j}{A})^2 = \sum_{j=1}^{p}((f_j - p_j) + \left(p_j - \frac{A_j}{A}\right))^2 = \sum_{j=1}^{p}(f_j - p_j)^2 - 2\sum_{j=1}^{p}\left(f_j - p_j\right)(p_j - \frac{A_j}{A}) + \sum_{j=1}^{p}(p_j - \frac{A_j}{A})^2$$

Prenons l'espérance mathématique de cette expression, il vient, compte tenu de $E(f_j\text{-}p_j) = 0$:

$$E(E_{R/A}^{*2}) = \sum_{j=1}^{p} E\left(\left(f_j - p_j\right)^2\right) + E_{R/A}^2$$

Or $E\left(\left(f_j - p_j\right)^2\right) = V(f_j) = \frac{p_j(1-p_j)}{n}$. Il en résulte que : $E\left(E_{R/A}^{*2}\right) = \frac{1-\sum p_j^2}{n} + E_{R/A}^2 < E_{R/A}^2 + \frac{1}{n}\#\, E_{R/A}^2$.
Le calcul de la variance est beaucoup plus lourd.