CHAPITRE 1-5
MESURE DE LA DISSIMILARITÉ

Plan

CHAPITRE 1-5
MESURE DE LA DISSIMILARITÉ

1-5.1 Multidimensionnalité, dissimilarité et concentration

PROBLÉMATIQUE DE LA MESURE DE LA DISSIMILARITÉ

Nous avons vu qu'une mesure associée à un concept établit une correspondance entre les objets et des nombres, ce qui permet de comparer les objets et de déterminer la valeur de vérité d'une ou de plusieurs des relations $=$, $\neq$, $>$ ou $<$. Si, comme cela arrive souvent, un concept comprend plusieurs dimensions, et que l'on veut néanmoins le traiter comme un tout, nous avons vu qu'il faut surmonter le problème de la multidimensionnalité et qu'on peut le faire en construisant un indice.

Mais il arrive que l'on soit confronté à des concepts auxquels on ne peut pas associer de mesure autre que catégorique. Impossible, alors, d'envisager la construction d'un indice. Considérons par exemple le concept de structure économique d'une ville ou d'une région. On a beau se satisfaire de définir la structure économique comme la répartition de l'emploi entre les branches d'activité, on ne peut guère associer à ce concept d'autre mesure qu'une classification (variable catégorique) : ville mono-industrielle, ville de services, etc. Mais comment arrive-t-on à construire une classification qui permette de bien saisir la réalité ? Une manière de procéder consiste à comparer les objets (en l'occurrence, les structures économiques observées) pour constituer des groupes d'objets assez similaires entre eux, et nettement différents des objets des autres groupes. Une telle classification peut ensuite servir de base à l'élaboration d'une typologie et à la définition d'une variable catégorique associée au concept.

Mentionnons en passant que, même lorsque la construction d'un indice est possible en principe, l'approche qui vient d'être évoquée peut être souhaitable s'il s'avère impossible de construire un indice qui soit satisfaisant au plan théorique. Ne pourrait-on pas, par exemple, étudier le développement humain en constituant une typologie des pays ? Une telle typologie permettrait de définir un indice approprié à chaque type de pays, de manière à ne comparer que des pays comparables, et avec des mesures adaptées aux caractéristiques de ces pays (c'est ce que fait déjà le PNUD par rapport à la mesure de la pauvreté : il calcule deux « Indices de la pauvreté humaine », l'un pour les pays en développement et l'autre, pour les pays développés).

La démarche qui consiste à classer les objets pour dégager des types ne peut qu'être grandement facilitée si l'on peut formaliser le concept de similarité et lui associer une mesure. Il existe d'ailleurs des procédures de classification automatique fondées sur des mesures de similarité [1]. En outre, on souhaitera parfois s'en tenir à une démarche heuristique, plus informelle, et examiner le degré de similarité entre des objets sans aller jusqu'à construire une typologie. Là encore, une mesure de la similarité peut être un outil précieux. C'est donc de la mesure de la similarité qu'il est question ici.

Notons d'abord que le concept de la similarité s'applique à une *paire* d'objets. La similarité n'est donc une propriété d'aucun des deux objets : elle est une propriété de la paire [2]. Ensuite, le concept de similarité est un concept général, qui recouvre une myriade de concepts spécifiques : car lorsqu'on examine la similarité entre deux objets, c'est toujours *par rapport* à un attribut donné. Un concept de similarité spécifique est défini par l'attribut auquel on se réfère pour comparer les objets dont on veut mesurer la similarité. S'agissant de villes, par exemple, on peut considérer la similarité par rapport à la structure démographique, par rapport au taux de criminalité, par rapport à la qualité de vie, etc.

Convenons d'emblée que la mesure de la similarité par rapport à un attribut unidimensionnel est une affaire triviale : il n'y a pas de problème particulier à mesurer, par exemple, la similarité entre deux pays quant au chiffre de leur population, au taux de criminalité ou à la valeur de l'IDH du PNUD [3]. Par contre, lorsqu'on veut mesurer la similarité par rapport à une propriété multidimensionnelle que l'on n'a pas au préalable résumée en un indice [4], on est confronté au même problème que dans la construction d'un nombre indice. Par exemple,

- Par rapport à leur structure économique, quel est le degré de similarité entre le Québec et l'Ontario ?
- Par rapport à leur répartition sur le territoire, quel est le degré de similarité entre la culture bananière et l'élevage au Costa Rica ?

---

1 Dendrogrammes, algorithmes de partition automatique, etc. Voir Legendre et Legendre (1984 et 1998).

2 On pourrait dire que l'objet auquel s'applique la similarité est une paire d'objets.

3 Cet exemple est délibérément paradoxal : alors que l'IDH est un indicateur qui cherche à mesurer une réalité multidimensionnelle, la comparaison de deux pays quant à la valeur de cet indicateur est, elle, unidimensionnelle.

4 Ou, ce qui revient au même, mesurer simultanément la similarité sous plusieurs aspects ou, pour le dire de façon elliptique, mesurer la similarité entre deux objets multidimensionnels.

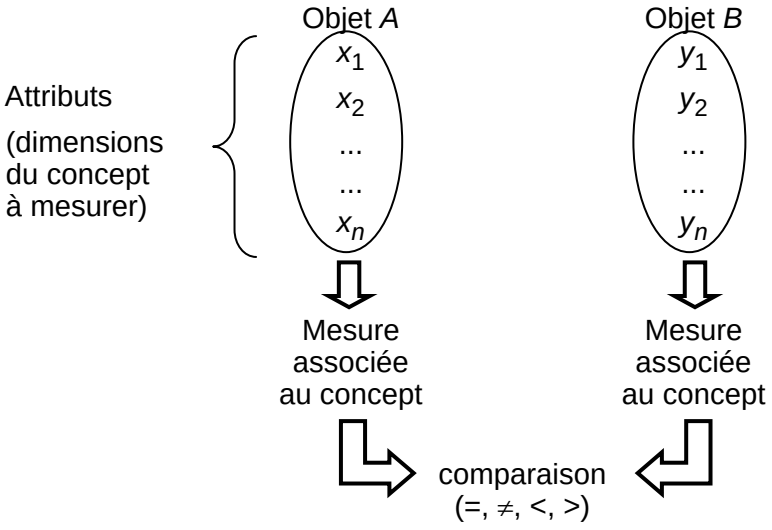
Pour mesurer la similarité dans les exemples qui précèdent, on doit tenir compte de plus d'une dimension, parce que le rapport sous lequel on examine la similarité réfère à un concept qui comprend plus d'une dimension :

- S'agissant de la similarité entre pays quant à leur structure économique, il faut tenir compte des différentes branches de la production.
- S'agissant de la similarité entre activités quant à la répartition spatiale, il faut tenir compte des différentes parties du territoire (zones, districts, provinces, ou autres, selon le découpage géographique utilisé).

En un sens, donc, les mesures de la similarité entre objets multidimensionnels s'apparentent à des indices. Pour bien expliciter les différences, nous allons nous attacher dans les lignes qui suivent à faire ressortir ce qu'il y a de spécifique à la mesure de la similarité.

Un indice, comme nous l'avons vu, résume en un seul chiffre les valeurs des indicateurs associés aux multiples dimensions d'un concept. L'indice est une mesure, parce qu'il permet de comparer deux objets quant au degré auquel ils possèdent la propriété définie par le concept. Cela est résumé dans le schémas suivant.

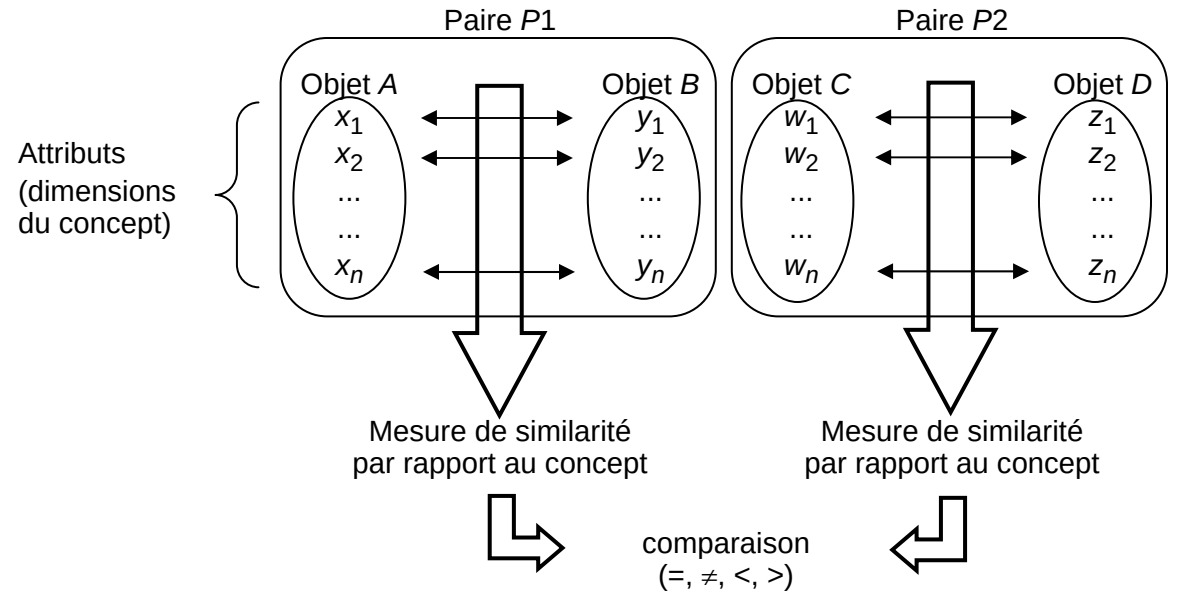
Construction d'un nombre indice



Pour mesurer la similarité, en revanche, on compare d'abord deux objets trait pour trait. On obtient ainsi autant de mesures partielles de similarité qu'il y a de dimensions à la comparaison. Il faut ensuite agréger toutes ces mesures partielles en une seule. Le résultat est une mesure

de la similarité : cette mesure permet de comparer deux *paires* d'objets quant à leur similarité par rapport à un attribut multidimensionnel donné. Cela est résumé dans le schéma qui suit. La comparaison entre les deux schémas montre en quoi la mesure de la similarité entre objets multidimensionnels est différente de la construction d'un indice.

Mesure de la similarité



Néanmoins, il s'agit bel et bien d'une mesure au sens où nous l'avons définie au chapitre 1-1. Rappelons qu'une mesure associée à un concept établit une correspondance entre les objets et des nombres, ce qui permet de comparer les objets et de déterminer la valeur de vérité d'une ou de plusieurs des relations $=$, $\neq$, $>$ ou $<$. Une mesure de la similarité est donc une correspondance qui permet de comparer deux *paires* d'objets quelconques du point de vue de leur similarité quant à un attribut donné. Formellement, si l'on convient que $f(A,B)$ est la mesure de la similarité entre les objets de la paire $[A,B]$ et que $f(C,D)$ est la mesure de la similarité entre les objets de la paire $[C,D]$, alors, une mesure de similarité permet de décider d'une ou de plusieurs des relations suivantes :

$f(A,B) = f(C,D)$

$f(A,B) \neq f(C,D)$

$f(A,B) < f(C,D)$

$f(A,B) > f(C,D)$

Par exemple, si *A* est le Nicaragua, *B* est le Costa Rica, *C* est le Costa Rica et *D* est le Canada [5], une mesure de similarité permet de répondre à la question « Par rapport à la composition de sa production, le Costa Rica ressemble-t-il davantage au Nicaragua ou au Canada ? ». De même, si *A* est la culture bananière, *B* est l'élevage, *C* est la culture bananière et *D* est la culture des agrumes, une mesure de la similarité permet de répondre à la question « Par rapport à sa répartition géographique au Costa Rica, la culture bananière ressemble-t-elle davantage à l'élevage ou à la culture des agrumes ? ».

Il est à noter que rien de ce qui précède n'implique que l'on mesure toujours la similarité selon une échelle rationnelle. Il est vrai que les variables utilisées comme mesures de similarité ou de dissimilarité sont souvent des variables rationnelles. Mais le problème de la multidimensionnalité fait qu'en général, il y a plusieurs mesures possibles et il n'y en a aucune qui puisse être considérée d'emblée comme la meilleure. C'est pourquoi, sauf dans des contextes particuliers, les mesures de similarité doivent normalement être interprétées comme des mesures ordinales : il faut se garder de leur donner une interprétation abusive de mesure d'intervalle ou rationnelle.

En outre, les mesures de similarité, comme nous le verrons, sont le plus souvent des *mesures inverses*, c'est-à-dire qu'elles sont en fait des mesures de *dissimilarité*. Il faut y être attentif, car cela peut causer la confusion.

LA MESURE DE LA DISSIMILARITÉ ENTRE DES DISTRIBUTIONS

Une distribution, ou une répartition, est une propriété (multidimensionnelle) d'une population (au sens général de collection de personnes ou d'objets), lorsque cette population est classée en catégories : c'est le nombre d'individus ou la fraction de la population qui se trouve dans chacune des catégories. Dans les exemples déjà évoqués,

- Les personnes employées dans une économie constituent une « population », que l'on peut classer entre les « catégories » que sont les branches d'activité. La structure économique du pays peut être décrite par une distribution : le nombre de personnes employées par branche d'activités.
- Les hectares de terre consacrés à une activité donnée (la culture bananière, par exemple) constituent une « population », que l'on peut classer entre les « catégories » que sont les

---

5 Comme le montre cet exemple, il peut arriver que $B=C$ (ou $B=D$, ou $A=C$, ou $A=D$), mais ce n'est pas nécessairement le cas.

subdivisions (provinces ou autres) d'un territoire. La répartition spatiale de l'activité peut être décrite par une distribution : le nombre d'hectares qui lui sont consacrés dans chaque subdivision du territoire.

Une distribution est donc un objet multidimensionnel. Mais la comparaison entre les distributions est grandement facilitée du fait qu'il existe une « règle de normalisation » naturelle : la mesure associée à chacune des dimensions de la distribution est simplement la fraction de la population appartenant à la catégorie correspondante. Or dans une distribution, la somme des parts est nécessairement égale à 1. Cela élimine d'emblée une partie du problème de la multidimensionnalité, celui, déjà mentionné à propos des nombres indices, du poids à accorder à chacune des dimensions.

Par contre, lorsqu'on tente de comparer deux objets qui ne sont pas des distributions, le choix de l'unité de mesure de chaque dimension de la comparaison détermine implicitement quel sera son poids dans la mesure de dissimilarité. Se pose alors dans toute son intensité le problème de multidimensionnalité évoqué à propos des nombres indices.

DISSIMILARITÉ ET INÉGALITÉ-CONCENTRATION : QUELLE DIFFÉRENCE ?

Dans les exemples donnés jusqu'ici, il s'est agi simplement d'examiner le degré d'association entre deux phénomènes ou inversement, le degré de ségrégation entre eux. Mais il y a une autre utilisation des mesures de dissimilarité entre deux distributions : c'est la mesure de la concentration ou de la dispersion. Une mesure de dissimilarité devient une mesure de concentration lorsqu'on compare la distribution étudiée avec une distribution de référence ou *théorique*. Cette distribution théorique, qui sert de point de référence, représente une concentration nulle et elle sert en quelque sorte d'étalon de mesure (nous verrons un exemple de cela plus loin).

Cela est cohérent avec ce que nous avons vu au chapitre 1-4 : en général, une mesure de l'inégalité compare la distribution observée avec une distribution de référence, qui représente l'égalité parfaite. Une mesure d'inégalité est donc une mesure de dissimilarité entre la distribution observée et la distribution de référence.

Il s'ensuit que l'indice de Gini est tout aussi approprié comme mesure de dissimilarité que comme mesure d'inégalité. D'ailleurs, nous avons déjà signalé parmi les propriétés de l'indice de Gini que celui-ci est symétrique, c'est-à-dire que les rôles de la distribution examinée et de la

distribution de référence sont interchangeables ; en d'autres mots, si l'on intervertit les rôles, la valeur du coefficient Gini est inchangée.

1-5.2 L'indice de dissimilarité

UN EXEMPLE NUMÉRIQUE

Nous considérons maintenant une mesure de dissimilarité largement utilisée, qui s'applique aux distributions comme, par exemple, la répartition géographique de l'emploi. Voici un exemple numérique fictif :

Emploi par zone et par branche

BRANCHE	*B*1	*B*2	*B*3	Total
ZONE				
*Z*1	48	325	287	660
*Z*2	27	185	148	360
*Z*3	45	90	45	180
Total	120	600	480	1200

Il s'agit de mesurer la similarité entre les branches d'activité quant à leur répartition géographique. On s'intéresse donc à la fraction de l'emploi de chaque branche dans chaque zone :

Distribution de l'emploi entre zones

BRANCHE	*B*1	*B*2	*B*3	Total
ZONE				
*Z*1	0,400	0,542	0,598	0,550
*Z*2	0,225	0,308	0,308	0,300
*Z*3	0,375	0,150	0,094	0,150
Total	1,000	1,000	1,000	1,000

La façon la plus simple qui se puisse imaginer d'examiner la similarité entre deux distributions consiste à regarder les différences entre ces fractions zone par zone. Faisons la comparaison entre les branches *B*1 et *B*2 :

Comparaison de la répartition géographique

des branches B1 et B2

BRANCHE	*B*1	*B*2	Écart
ZONE			
*Z*1	0,400	0,542	0,142
*Z*2	0,225	0,308	0,083
*Z*3	0,375	0,150	–0,225
Total	1,000	1,000	0,000

Chacun des écarts calculés constitue l'une des dimensions de la dissimilarité entre les deux répartitions géographiques. Pour mesurer la dissimilarité, il faut combiner les écarts en un chiffre unique. Il va de soi qu'une simple addition donnera toujours le même résultat, zéro [6]. C'est pourquoi l'on fera la somme des valeurs absolues :

$$|0{,}142| + |0{,}083| + |-0{,}225| = 0{,}142 + 0{,}083 + 0{,}225 \text{ (et non pas } -0{,}225)$$

Pour des raisons qui deviendront évidentes plus loin, on divise le résultat par deux et on obtient :

$$\frac{|0{,}142| + |0{,}083| + |-0{,}225|}{2} = 0{,}225$$

DÉFINITION DE L'INDICE DE DISSIMILARITÉ

La mesure de la dissimilarité dans une table de contingence : rappel de la notation

Pour formaliser la présentation, nous reprenons, en la généralisant, la notation développée à la section 1-2.1 [7]. Nous traitons une table de contingence à deux dimensions. Convenons que les colonnes correspondent à *n groupes* différents, alors que les lignes correspondent à *m catégories* différentes (dans notre exemple, comme à la section 1-2.1, les *n* « groupes » sont les 3 branches d'activité, tandis que les *m* « catégories » sont les 3 zones.

---

6 Puisque $\sum_i v_i = \sum_i w_i = 1$, alors $\sum_i (v_i - w_i) = \sum_i v_i - \sum_i w_i = 0$.

7 Le lecteur est invité à se référer à la section 1-2.1 pour un énoncé des identités fondamentales qui se vérifient dans une table de contingence.

x_{ij}	nombre d'emplois de la branche *j* dans la zone *i*
$x_{\bullet j} = \sum_i x_{ij}$	nombre total d'emplois de la branche *j*
$x_{i\bullet} = \sum_j x_{ij}$	nombre total d'emplois dans la zone *i*
$x_{\bullet\bullet} = \sum_i \sum_i x_{ij}$	nombre total d'emplois de toutes branches dans toutes zones
$p_{ij} = x_{ij} / x_{\bullet\bullet}$	fraction de l'emploi total global qui appartient à la branche *j* **et** est situé dans la zone *i*
$p_{\bullet j} = \sum_i p_{ij}$	fraction de l'emploi total global qui appartient à la branche *j*
$p_{i\bullet} = \sum_j p_{ij}$	fraction de l'emploi total global qui est situé dans la zone *i*
$p_{j/i\bullet} = p_{ij} / p_{i\bullet}$	fraction de l'emploi total de la zone *i* qui appartient à la branche *j*
$p_{i/\bullet j} = p_{ij} / p_{\bullet j}$	fraction de l'emploi total de la branche *j* qui est situé dans la zone *i*

Dans l'exemple numérique ci-haut, nous avons appliqué une mesure de dissimilarité entre deux répartitions géographiques, celle de la branche *B*1 et celle de la branche *B*2. Selon la notation courante, cela correspond à l'application d'une mesure de dissimilarité aux distributions données par les vecteurs

$$Q_1 = \begin{bmatrix} p_{1/\bullet 1} \\ p_{2/\bullet 1} \\ \vdots \\ p_{m/\bullet 1} \end{bmatrix} \text{ et } Q_2 = \begin{bmatrix} p_{1/\bullet 2} \\ p_{2/\bullet 2} \\ \vdots \\ p_{m/\bullet 2} \end{bmatrix}$$

Plus généralement, on compare les distributions

$$Q_h = \begin{bmatrix} p_{1/\bullet h} \\ p_{2/\bullet h} \\ \vdots \\ p_{m/\bullet h} \end{bmatrix} \text{ et } Q_k = \begin{bmatrix} p_{1/\bullet k} \\ p_{2/\bullet k} \\ \vdots \\ p_{m/\bullet k} \end{bmatrix}$$

ou encore les distributions

$$R_g = \begin{bmatrix} p_{1/g\bullet} & p_{2/g\bullet} & \cdots & p_{n/g\bullet} \end{bmatrix} \text{ et } R_i = \begin{bmatrix} p_{1/i\bullet} & p_{2/i\bullet} & \cdots & p_{n/i\bullet} \end{bmatrix}$$

NOTE : On peut travailler soit avec des fractions, comme dans la notation ci-haut, soit avec des pourcentages, obtenus en multipliant les fractions par 100. Ici, nous convenons de travailler avec des fractions, parce que cela allège l'écriture des formules. Mais la pratique courante dans la présentation des résultats consiste à présenter des pourcentages, ce qui allège les tableaux, grâce à l'élimination de la virgule décimale.

Définition

Dans ce qui suit, nous appliquons l'indice de dissimilarité à une comparaison des distributions Q_h et Q_k. Tous les développements peuvent se transposer aisément à une comparaison entre les distributions R_g et R_i ou, en vérité, à n'importe quelle paire de distributions formellement comparables (c'est-à-dire ayant le même nombre de possibilités).

L'indice de dissimilarité se définit comme

$$D = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right|$$

Dans l'exemple numérique donné ci-haut,

$$Q_1 = \begin{bmatrix} 0{,}400 \\ 0{,}225 \\ 0{,}375 \end{bmatrix} \text{ et } Q_2 = \begin{bmatrix} 0{,}542 \\ 0{,}308 \\ 0{,}150 \end{bmatrix}$$

et

$$D = \frac{\left| 0{,}400 - 0{,}542 \right| + \left| 0{,}225 - 0{,}308 \right| + \left| 0{,}375 - 0{,}150 \right|}{2} = 0{,}225$$

Cet indice de dissimilarité et ses proches variantes apparaissent sous divers noms dans différentes disciplines. Par exemple,

- On désigne aussi l'indice de dissimilarité par les expressions « indice de différenciation » et « indicateur de dissociation ».
- Lorsque l'une des distributions est la répartition spatiale d'une activité économique et l'autre, celle de l'*ensemble* des activités, cette mesure correspond à que l'on appelle en science régionale le *coefficient de localisation*. Il est à noter toutefois que le coefficient de localisation, bien qu'il se calcule au moyen de la même formule, n'est pas à proprement parler un indice de dissimilarité : nous verrons pourquoi plus loin.

- On connaît aussi, en sciences régionales, le coefficient de *spécialisation*, qui compare la structure économique d'une zone (répartition de l'emploi entre les branches d'activité) avec celle de l'ensemble du territoire à l'étude. Cette mesure non plus, n'est pas à proprement parler un indice de dissimilarité.
- En géographie, Taylor (1977, p. 180) cite une multitude d'appellations pour l'indice de dissimilarité ; parmi ces appellations, « coefficient d'association géographique » est particulièrement déroutante, puisque *D* est une mesure de *dissimilarité*, ou de *dissociation*. On trouve même des géographes qui utilisent le terme « indice de Gini » pour désigner l'indice de dissimilarité...
- Les démographes et les sociologues utilisent ce même indice, sous le nom de « coefficient de ségrégation résidentielle » ou « indice de discrimination », pour comparer les distributions spatiales résidentielles de différents groupes ethniques ou raciaux (Mills et Hamilton, 1989, p.233-239 ; Waldorf, 1993).

Que faut-il retenir de cette confusion terminologique ? Ceci : lorsque vous prenez connaissance de résultats de recherche qui font appel à des indices de ce type, assurez-vous de bien vérifier quelle est la formule mathématique utilisée.

Au delà des particularités propres à chaque discipline, examinons cet indice de dissimilarité en tant que mesure de dissimilarité entre deux distributions.

L'INDICE DE DISSIMILARITÉ COMME MESURE DE CONCENTRATION OU D'INÉGALITÉ

Jusqu'à maintenant, nous avons discuté des utilisations de l'indice de dissimilarité pour mesurer la dissimilarité entre deux distributions observées. Mais on peut aussi utiliser l'indice de dissimilarité pour mesurer l'inégalité ou la concentration. D'ailleurs, répétons-le, les mesures d'inégalité ou de concentration sont généralement des mesures de dissimilarité entre une distribution observée et une distribution de référence. Pour mesurer l'inégalité ou la concentration, il s'agit donc de comparer une distribution observée avec une distribution de référence, qui représente l'égalité parfaite ou une concentration nulle (évidemment, dans ce cas, le tableau des données n'est pas une table de contingence).

Exemple

Supposons que l'on veuille mesurer le degré de concentration géographique de la population sur un territoire donné, préalablement découpé en zones (provinces, districts, ...). Une

concentration nulle correspond à une situation où la densité de la population (habitants/km^2) est partout la même. On peut donc dire que la concentration est nulle si la fraction de la population dans chaque zone est égale à la fraction du territoire compris dans cette zone.

Soit V, la distribution de la superficie du territoire et W, celle de la population.

$$V = \lfloor v_1 \quad v_2 \quad \cdots \quad v_n \rfloor \text{ et } W = \lfloor w_1 \quad w_2 \quad \cdots \quad w_n \rfloor$$

v_i est la fraction de la superficie totale qui est comprise dans la zone i et w_i est la fraction de la population qui se trouve dans la zone i.

La concentration est nulle si

$$w_i = v_i \text{ pour tout } i$$

Dans ce cas, la distribution *observée* du territoire sert de distribution *de référence* à la population : elle est la distribution théorique ou hypothétique d'une population de concentration nulle [8]. On peut alors utiliser l'indice de dissimilarité entre la distribution du territoire et la distribution de la population comme mesure de la concentration géographique de la population. On aura

$$D = \frac{1}{2} \sum_i \left| w_i - v_i \right|$$

Le tableau qui suit illustre cette utilisation de l'indice de dissimilarité. On y mesure le degré de concentration de la population de la Ville de Montréal. Les données de population sont celles du Recensement de 1991. Le territoire est découpé selon les 54 quartiers de planification de la Ville, rangés par ordre décroissant de densité. On obtient $D = 0{,}2361$, c'est-à-dire que, pour obtenir une densité uniforme, il faudrait déplacer 23,61 % de la population d'un quartier à un autre.

---

8 En d'autres mots, la distribution V est observée quand il s'agit du territoire, mais elle devient hypothétique quand on l'applique à la population

Mesure de la concentration de la population au moyen de l'indice de dissimilarité :

Ville de Montréal (54 quartiers de planification), population Recensement 1991

Quartier	Données Pop. 1991	Superf. km^2	Densité hab/km^2	Répartitions Pop.	Superf.	Écart absolu
11	29469	1,65	17860	2,90%	0,88%	0,0201
8	10604	0,72	14728	1,04%	0,38%	0,0066
18	27022	2,03	13311	2,66%	1,08%	0,0157
34	24258	1,85	13112	2,38%	0,99%	0,0140
13	30314	2,39	12684	2,98%	1,28%	0,0170
35	14187	1,24	11441	1,39%	0,66%	0,0073
31	19652	1,73	11360	1,93%	0,92%	0,0101
33	15752	1,40	11251	1,55%	0,75%	0,0080
42	25495	2,32	10989	2,51%	1,24%	0,0127
15	19126	1,75	10929	1,88%	0,93%	0,0095
16	15030	1,38	10891	1,48%	0,74%	0,0074
29	15606	1,46	10689	1,53%	0,78%	0,0075
9	21348	2,02	10568	2,10%	1,08%	0,0102
32	14737	1,48	9957	1,45%	0,79%	0,0066
40	20350	2,15	9465	2,00%	1,15%	0,0085
14	15973	1,80	8874	1,57%	0,96%	0,0061
10	14165	1,65	8585	1,39%	0,88%	0,0051
27	11592	1,41	8221	1,14%	0,75%	0,0039
17	16167	2,00	8084	1,59%	1,07%	0,0052
30	29664	3,69	8039	2,91%	1,97%	0,0095
45	24738	3,23	7659	2,43%	1,72%	0,0071
46	19880	2,60	7646	1,95%	1,39%	0,0057
39	34906	4,85	7197	3,43%	2,59%	0,0084
51	8452	1,20	7043	0,83%	0,64%	0,0019
23	18672	2,67	6993	1,83%	1,43%	0,0041
12	14980	2,21	6778	1,47%	1,18%	0,0029
6	16785	2,48	6768	1,65%	1,32%	0,0033
19	11499	1,75	6571	1,13%	0,93%	0,0020
4	23636	3,70	6388	2,32%	1,98%	0,0035
44	18699	2,96	6317	1,84%	1,58%	0,0026
24	13665	2,22	6155	1,34%	1,19%	0,0016
21	20564	3,62	5681	2,02%	1,93%	0,0009
48	17038	3,02	5642	1,67%	1,61%	0,0006
41	20092	3,59	5597	1,97%	1,92%	0,0006
5	18478	3,36	5499	1,82%	1,79%	0,0002
49	14687	2,73	5380	1,44%	1,46%	0,0001
20	27819	5,22	5329	2,73%	2,79%	0,0005
43	24957	4,84	5156	2,45%	2,58%	0,0013
3	18052	3,56	5071	1,77%	1,90%	0,0013
28	17764	3,56	4990	1,75%	1,90%	0,0015
2	25181	5,25	4796	2,47%	2,80%	0,0033
26	19073	4,01	4756	1,87%	2,14%	0,0027
22	9651	2,18	4427	0,95%	1,16%	0,0022
38	12512	3,16	3959	1,23%	1,69%	0,0046
7	22660	5,84	3880	2,23%	3,12%	0,0089
1	22613	5,85	3865	2,22%	3,12%	0,0090
52	35098	9,50	3695	3,45%	5,07%	0,0162
50	14403	4,07	3539	1,42%	2,17%	0,0076
47	13111	4,45	2946	1,29%	2,38%	0,0109
54	47534	19,04	2497	4,67%	10,16%	0,0549
37	3546	2,06	1721	0,35%	1,10%	0,0075
25	4009	4,28	937	0,39%	2,28%	0,0189
53	11970	13,92	860	1,18%	7,43%	0,0625
36	431	4,24	102	0,04%	2,26%	0,0222
Total	1017666	187,34	5432	100,00%	100,00%	0,472

Indice de dissimilarité : 0,2361

PROPRIÉTÉS DE L'INDICE DE DISSIMILARITÉ

L'indice de dissimilarité et les propriétés d'une mesure d'inégalité

Puisqu'une mesure d'inégalité est une mesure de dissimilarité entre la distribution observée et une distribution de référence, les propriétés désirables d'une mesure d'inégalité sont également désirables d'une mesure de dissimilarité. Qu'en est-il donc de l'indice de dissimilarité *D* ?

Rappelons les propriétés désirables d'une mesure d'inégalité selon Valeyre (1993) :

1. Une mesure d'inégalité doit prendre des valeurs non négatives, puisqu'il s'agit d'une mesure de l'éloignement de la distribution observée par rapport à la distribution de référence.
2. Une mesure d'inégalité doit prendre la valeur zéro si, et seulement si, la distribution observée est identique à la distribution de référence.
3. Toutes les observations doivent être traitées de la même manière.
4. Une mesure d'inégalité doit être indépendante de la valeur moyenne de la variable examinée ; une mesure de concentration doit être indépendante de la taille de la population dont on étudie la distribution.
5. L'agrégation d'observations affichant le même degré de spécificité ne doit pas changer la valeur de la mesure.
6. Une mesure d'inégalité doit diminuer si la distribution est modifiée d'une façon qui réduit incontestablement l'inégalité (Principe de transfert de Pigou-Dalton).

L'indice de dissimilarité possède les propriétés 1 à 5, mais pas la propriété 6 : sa valeur demeure inchangée après un transfert entre deux catégories dont les spécificités sont toutes deux supérieures ou toutes deux inférieures à 1 [9].

Domaine de variation

Si l'on vous dit que vous avez obtenu la note 18 à un examen, serez-vous content ? Cette note est-elle une bonne, ou une mauvaise note ? Pour le savoir, il faut d'abord savoir quelle est la

---

[9] On peut démontrer cette caractéristique de façon très simple à l'aide de l'interprétation géométrique de l'indice de dissimilarité comme distance verticale maximum entre la courbe de Lorenz et la diagonale. Voir ci-après.

note maximale [10]. Si l'examen est noté sur 20, la note 18 est probablement une bonne note ; s'il est noté sur 100, vous ne serez sans doute pas content...

C'est pour cette raison que l'on s'intéresse au domaine de variation d'une mesure. Le domaine de variation d'une mesure est l'ensemble des valeurs qu'elle peut prendre. Pour une mesure continue, le domaine de variation est défini par la valeur minimum et la valeur maximum que peut prendre la mesure. Pour pouvoir savoir si une valeur donnée est « grande » ou non, il faut au moins connaître son domaine de variation, pour voir si cette valeur est plus proche du maximum ou du minimum.

Dans le cas de l'indice de dissimilarité, sa valeur minimum est zéro : cet indice prend la valeur zéro quand $p_{i/\bullet h} = p_{i/\bullet k}$ pour tout *i*, c'est-à-dire quand les distributions sont identiques.

Quelle est sa valeur maximum ?

Lorsque l'on compare les distributions de deux populations parfaitement distinctes [11], la valeur maximum que peut prendre l'indice est 1 : cela se produit quand $p_{i/\bullet h} = 0$ lorsque $p_{i/\bullet k} > 0$ et vice-versa, c'est-à-dire quand la séparation entre les deux populations est complète : elles ne sont jamais présentes ensemble dans la même catégorie. Dans cette situation en effet, pour chaque catégorie *i*, on a

SOIT $p_{i/\bullet h} = 0$, et alors

$$\left|p_{i/\bullet h} - p_{i/\bullet k}\right| = \left|0 - p_{i/\bullet k}\right| = p_{i/\bullet k} = 0 + p_{i/\bullet k} = p_{i/\bullet h} + p_{i/\bullet k}$$

SOIT $p_{i/\bullet k} = 0$, et alors

$$\left|p_{i/\bullet h} - p_{i/\bullet k}\right| = \left|p_{i/\bullet h} - 0\right| = p_{i/\bullet h} = p_{i/\bullet h} + 0 = p_{i/\bullet h} + p_{i/\bullet k}$$

On a donc

$$D^{\max} = \frac{1}{2}\sum_i \left|p_{i/\bullet h} - p_{i/\bullet k}\right| = \frac{1}{2}\sum_i \left(p_{i/\bullet h} + p_{i/\bullet k}\right)$$

$$D^{\max} = \frac{1}{2}\left(\sum_i p_{i/\bullet h} + \sum_i p_{i/\bullet k}\right) = \frac{1+1}{2} = 1$$

---

[10] Ce n'est pas la seule considération. L'interprétation de la note dépend aussi de la note obtenue par les autres et des critères qui sont communément utilisés pour l'interpréter (comme la note de passage).

La division par 2, dans la formule de calcul de l'indice de dissimilarité, a donc pour effet de « normaliser » son domaine de variation à l'intervalle [0, 1].

L'indice de dissimilarité ne peut-il pas prendre une valeur supérieure à 1 ? Non. Pour s'en convaincre, il suffit de se demander, à partir de la situation de séparation complète décrite ci-haut, quelle serait la conséquence de déplacer un individu d'une catégorie à une autre (effet nul si cet individu reste avec ceux de son espèce ; autrement, la valeur de l'indicateur diminue). L'exemple numérique suivant illustre le cas de la ségrégation totale.

Indice de dissimilarité : exemple de ségrégation totale

	Nombres			Répartitions			
ETHNIE	Martiens	Terriens	Total	Martiens	Terriens	Total	Écart
	x_{i1}	x_{i2}	$x_{i1} + x_{i2}$	$p_{i/\bullet 1}$	$p_{i/\bullet 2}$	$p_{i\bullet}$	$\left\lvert p_{i/\bullet h} - p_{i/\bullet k} \right\rvert$
PLANÈTE							
TERRE	0	6	6	0,00	0,75	0,40	0,75
LUNE	0	2	2	0,00	0,25	0,13	0,25
MARS	3	0	3	0,43	0,00	0,20	0,43
JUPITER	4	0	4	0,57	0,00	0,27	0,57
TOTAL	7	8	15	1,00	1,00	1,00	

Indice de dissimilarité :

$$\frac{0{,}75 + 0{,}25 + 0{,}43 + 0{,}57}{2} = 1{,}00$$

Interprétation métaphorique

Même si l'on connaît parfaitement le domaine de variation d'une mesure, il est parfois difficile d'avoir une intuition concrète de ce qu'est une « grande » valeur. D'où l'utilité d'une interprétation métaphorique. Une interprétation métaphorique, comme son nom l'indique, repose sur une comparaison, une métaphore : « C'est comme si »... Il faut bien se garder de prendre ces interprétations métaphoriques au pied de la lettre.

Pour ce qui est de l'indice de dissimilarité, il compare la distribution de deux groupes parfaitement distincts [12], disons *h* et *k*. On peut interpréter l'indice comme la fraction du groupe *h* qu'il faudrait déplacer d'une catégorie à l'autre, pour que sa distribution soit identique à celle du groupe *k*.

---

11 On entend par là qu'aucun individu n'appartient aux deux populations à la fois.

12 On entend par là qu'aucun individu n'appartient aux deux populations à la fois.

Ainsi, dans l'exemple numérique donné au début de cette section, l'indice de dissimilarité entre la répartition spatiale des emplois de la branche *B*1 et ceux de la branche *B*2 est de 0,225. Cela signifie que, pour rendre la répartition spatiale de *B*1 identique à celle de *B*2, il faudrait déplacer 22,5 % des emplois de *B*1.

Ce résultat est facile à démontrer. Commençons par déterminer quelle est la fraction du groupe *h* qu'il faudrait déplacer pour passer de la distribution représentée par les $p_{i/\bullet h}$ à la distribution représentée par les $p_{i/\bullet k}$. Il suffit pour cela d'additionner les fractions de population à retirer des catégories (zones, régions,...) « excédentaires » pour les redistribuer dans des catégories « déficitaires ». Désignons par *A* l'ensemble des catégories « excédentaires », c'est-à-dire où $p_{i/\bullet h} > p_{i/\bullet k}$. Pour chacune des catégories appartenant à l'ensemble *A*, la fraction « excédentaire » de la population *h* est égale à $p_{i/\bullet h} - p_{i/\bullet k}$. Au total, la fraction de la population *h* à retirer des catégories « excédentaires » est donc donnée par

$$\sum_{i \in A} \left(p_{i/\bullet h} - p_{i/\bullet k} \right)$$

On peut de façon équivalente additionner les fractions de population à ajouter aux catégories « déficitaires », c'est-à-dire

$$\sum_{i \notin A} \left(p_{i/\bullet k} - p_{i/\bullet h} \right)$$

Naturellement, $\sum_{i \in A} \left(p_{i/\bullet h} - p_{i/\bullet k} \right) = \sum_{i \notin A} \left(p_{i/\bullet k} - p_{i/\bullet h} \right)$

puisque $\sum_{i \in A} \left(p_{i/\bullet h} - p_{i/\bullet k} \right) - \sum_{i \notin A} \left(p_{i/\bullet k} - p_{i/\bullet h} \right)$

$$= \sum_{i \in A} p_{i/\bullet h} - \sum_{i \in A} p_{i/\bullet k} - \sum_{i \notin A} p_{i/\bullet k} + \sum_{i \notin A} p_{i/\bullet h} = \sum_{i} p_{i/\bullet h} - \sum_{i} p_{i/\bullet k} = 0$$

Quel rapport avec l'indice de dissimilarité ? Eh bien, si l'on additionne les deux sommations du membre de gauche de l'équation précédente (au lieu de soustraire la seconde de la première), on obtient

$$\sum_{i \in A} \left(p_{i/\bullet h} - p_{i/\bullet k} \right) + \sum_{i \notin A} \left(p_{i/\bullet k} - p_{i/\bullet h} \right) = \sum_{i} \left| p_{i/\bullet h} - p_{i/\bullet k} \right|$$

Et puisque les deux termes du membre de droite sont égaux, on a donc

$$\sum_{i \in A} \left(p_{i/\bullet h} - p_{i/\bullet k} \right) = \sum_{i \notin A} \left(p_{i/\bullet k} - p_{i/\bullet h} \right) = \frac{1}{2} \sum_{i} \left| p_{i/\bullet h} - p_{i/\bullet k} \right| = D$$

Et voilà une autre bonne raison de diviser la somme par 2 !

Symétrie

Il est à noter que l'indice de dissimilarité *D* est symétrique par rapport aux groupes *h* et *k* :

$$D = \frac{1}{2} \sum_{i} \left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \frac{1}{2} \sum_{i} \left| p_{i/\bullet k} - p_{i/\bullet h} \right|$$

Par conséquent, on peut tout aussi bien interpréter l'indicateur comme la fraction du groupe *k* qu'il faudrait déplacer pour que sa distribution soit identique à celle du groupe *h* : quel que soit le groupe que l'on imagine de déplacer pour rendre sa distribution identique à celle de l'autre, la fraction à déplacer est la même (ainsi, on pourrait dire qu'il faudrait déplacer 22,5 % de l'emploi de la branche *B*2 pour rendre sa distribution identique à celle de *B*1). Pour ce qui est toutefois du *nombre* d'individus à déplacer, il est naturellement égal à cette fraction, multipliée par le chiffre de la population. Si les deux groupes sont de taille différente, le nombre d'individus à déplacer (hypothétiquement) sera différent selon que l'on imagine de déplacer une fraction de l'un ou de l'autre.

Insistons de nouveau sur le caractère métaphorique de cette interprétation. D'abord, la similarité des distributions n'est pas nécessairement une bonne chose (qu'on pense à la controverse à propos du *busing* qui a été pratiqué aux États-Unis pour réaliser l'intégration scolaire des Blancs et des Noirs). Ensuite, le déplacement (forcé) des populations n'est décidément pas un moyen acceptable lorsqu'il s'agit de populations humaines.

Autres propriétés

L'indice de dissimilarité, comme tout indice, a ses limites. Outre le non-respect du principe de transfert de Pigou-Dalton, mentionnons :

- Quand les données sont groupées, l'indice de dissimilarité, comme l'indice de Gini, est sensible à la définition et au nombre des catégories utilisées (classes, zones). Cette faiblesse n'est pas trop grave si le découpage choisi est assez fin – s'il comprend un grand

nombre de catégories – mais les comparaisons entre découpages différents sont sans signification [13].

- Lorsqu'il est utilisé comme mesure de concentration spatiale, l'indice de dissimilarité, comme l'indice de Gini, ne tient aucun compte de la contiguïté ou de la proximité des unités spatiales.
- L'indice de dissimilarité n'admet pas de données négatives. Par exemple, on ne pourrait pas utiliser l'indice de dissimilarité pour mesurer la similarité entre deux branches d'activité quant aux *variations* du nombre d'emplois par zone, parce que ces variations peuvent être négatives.

L'indice de dissimilarité et la courbe de Lorenz

Nous venons de voir que, comme l'indice de Gini. l'indice de dissimilarité peut servir à mesurer la concentration, bien qu'il ne possède pas toutes les propriétés désirables de l'indice de Gini (il lui manque le principe de transfert de Pigou-Dalton). Nous avons vu aussi que l'indice de Gini peut se calculer géométriquement, à partir de la courbe de Lorenz. Existe-t-il un rapport entre l'indice de dissimilarité et la courbe de Lorenz ? Oui !

Il se trouve en effet que l'indice de dissimilarité est égal à l'écart vertical maximum entre la courbe de Lorenz et la diagonale

$$D = \underset{k}{\text{MAX}} \left| Cv_k - Cw_k \right|$$

Démonstration :

Puisque

$$\sum_i \left(v_i - w_i \right) = \sum_i v_i - \sum_i w_i = 1 - 1 = 0 \ ,$$

cette somme contient des termes positifs et des termes négatifs (à moins que tous les termes ne soient nuls). Or l'ordonnancement des observations en ordre croissant des rapports w_i / v_i fait en sorte que les termes $(v_i - w_i)$ qui sont positifs précèdent ceux qui sont négatifs. Alors, il est évident que l'écart vertical

[13] Cette question est discutée dans les écrits en géographie sous la rubrique MAUP, c'est-à-dire « Modifiable Areal Unit Problem ».

$$Cv_k - Cw_k = \sum_{i-1}^{k} v_i - \sum_{i-1}^{k} w_i = \sum_{i-1}^{k} (v_i - w_i)$$

atteint sa valeur maximum quand on choisit k de façon à inclure dans la sommation tous les termes positifs, tout en excluant tous ceux qui sont négatifs. Donc

$$\underset{k}{\text{MAX}} \left[Cv_k - Cw_k \right] = \sum_{\substack{i \text{ tel que} \\ v_i > w_i}} (v_i - w_i)$$

où, puisque $\sum_i (v_i - w_i) = 0$,

$$\sum_{\substack{i \text{ tel que} \\ v_i > w_i}} (v_i - w_i) = \sum_{\substack{i \text{ tel que} \\ v_i < w_i}} |v_i - w_i| = \frac{1}{2} \sum_i |v_i - w_i| = D$$

L'indice de dissimilarité D trouve ainsi une interprétation géométrique : c'est la distance maximum entre la diagonale et la courbe de Lorenz associée à la distribution V (voir l'exemple numérique tiré de Taylor, 1977, et discuté en 1-4.3).

Il est aisé de constater, à l'aide de cette interprétation, que l'indice de dissimilarité est insensible à toute redistribution qui ne réduit pas l'écart vertical maximum mais qui rapproche néanmoins la courbe de Lorenz de la diagonale. C'est cette insensibilité qui viole le principe de transfert de Pigou-Dalton.

Sommaire des propriétés de l'indice de dissimilarité

1. Possède les 5 premières propriétés désirables d'une mesure d'inégalité, mais pas la dernière (il manque le principe de transfert de Pigou-Dalton ; Valeyre, 1993)
2. Domaine de variation (valeurs maximum et minimum)
   - $D = 0$ quand $p_{i/\bullet h} = p_{i/\bullet k}$ pour tout i (les deux distributions sont identiques)
   - $D = 1$ quand il y a ségrégation complète :
     soit $p_{i/\bullet k} > 0$, et alors, $p_{i/\bullet h} = 0$
     soit $p_{i/\bullet h} > 0$, et alors, $p_{i/\bullet k} = 0$
3. D est symétrique par rapport aux groupes h et k :
   $$D = \frac{1}{2} \sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \frac{1}{2} \sum_i \left| p_{i/\bullet k} - p_{i/\bullet h} \right|$$
4. Interprétation métaphorique (groupes parfaitement distincts) :

D = fraction du groupe *h* qu'il faudrait déplacer pour que sa distribution soit identique à celle du groupe *k* ou vice-versa.

5. Quand les données sont groupées, *D*, aussi bien que *G*, est sensible à la définition et au nombre de catégories utilisées (classes, zones).

   Cela implique notamment que l'agrégation d'une ou de plusieurs catégories peut entraîner une diminution de la valeur de l'indice de dissimilarité.
6. En tant que mesure de concentration spatiale, l'indice de dissimilarité, comme le Gini, ne tient aucun compte de la proximité dans l'espace des différentes zones de forte densité.
7. Ne s'applique pas à des données négatives (ex. : comparaison des variations de l'emploi).
8. *D* est égal à l'écart vertical maximum entre la courbe de Lorenz et la diagonale.

APPLICATION DE L'INDICE DE DISSIMILARITÉ À UNE DICHOTOMIE

Équivalence de la formule de Duncan et Duncan (1955)

Lorsqu'on ne distingue que deux groupes, on a affaire à une dichotomie : on compare alors un groupe *h* avec le reste de la population (qui joue le rôle du groupe *k*). Pour le groupe *k*, on a alors

$$p_{i/\bullet k} = \frac{x_{i\bullet} - x_{ih}}{x_{\bullet\bullet} - x_{\bullet h}} = \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}}$$

Et dans ce cas, on peut écrire l'indice de dissimilarité sous la forme

$$D = \frac{\sum_i p_{i\bullet} \left| p_{h/i\bullet} - p_{\bullet h} \right|}{2\, p_{\bullet h} \left(1 - p_{\bullet h}\right)}$$

Cette seconde définition, qui est celle donnée dans l'article classique de Duncan et Duncan (1955), est équivalente à celle que nous avons donnée précédemment, lorsqu'on l'applique à une dichotomie.

L'équivalence entre les deux définitions dans le cas d'une dichotomie se démontre comme suit :

$$D = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}} \right| = \frac{1}{2}\sum_i \left| \frac{p_{ih}}{p_{\bullet h}} - \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}} \right|$$

$$D = \frac{1}{2}\sum_i p_{i\bullet}\left|\frac{p_{ih}/p_{i\bullet}}{p_{\bullet h}} - \frac{1-\left(p_{ih}/p_{i\bullet}\right)}{1-p_{\bullet h}}\right| = \frac{1}{2}\sum_i p_{i\bullet}\left|\frac{p_{h/i\bullet}}{p_{\bullet h}} - \frac{1-p_{h/i\bullet}}{1-p_{\bullet h}}\right|$$

$$D = \frac{\sum_i p_{i\bullet}\left|p_{h/i\bullet}\left(1-p_{\bullet h}\right) - \left(1-p_{h/i\bullet}\right)p_{\bullet h}\right|}{2\,p_{\bullet h}\left(1-p_{\bullet h}\right)}$$

$$D = \frac{\sum_i p_{i\bullet}\left|p_{h/i\bullet} - p_{h/i\bullet}\,p_{\bullet h} - p_{\bullet h} + p_{h/i\bullet}\,p_{\bullet h}\right|}{2\,p_{\bullet h}\left(1-p_{\bullet h}\right)} = \frac{\sum_i p_{i\bullet}\left|p_{h/i\bullet} - p_{\bullet h}\right|}{2\,p_{\bullet h}\left(1-p_{\bullet h}\right)}$$

Cette formule se prête à une interprétation intéressante.

Au numérateur, on a une moyenne pondérée des écarts absolus $\left|p_{h/i\bullet} - p_{\bullet h}\right|$ entre, d'une part, la proportion $p_{h/i\bullet}$ du groupe *h* dans chaque catégorie *i* et, d'autre part, la proportion $p_{\bullet h}$ du groupe *h* dans l'ensemble de la population : le poids $p_{i\bullet}$ de chaque catégorie est proportionnel à sa population, tous groupes confondus.

Quant à l'expression au dénominateur, elle est égale à l'écart absolu moyen entre les *individus* (et non pas entre les catégories) de la variable dichotomique d'appartenance au groupe *h*. Cet écart absolu moyen est égal à deux fois la variance de la même variable.

Soit en effet la variable dichotomique d'appartenance g_t :

$$g_t \begin{cases} = 1 \text{ si l'individu } t \text{ appartient au groupe } h \\ = 0 \text{ autrement} \end{cases}$$

où l'indice *t* se rapporte aux individus des deux groupes : *t* varie de 1 à $x_{\bullet\bullet}$.

La variable g_t a une distribution binomiale, dont la moyenne est donnée par

$$\mu_g = \frac{\sum_t g_t}{x_{\bullet\bullet}} = \frac{\sum_i x_{ih}}{x_{\bullet\bullet}} = \frac{x_{\bullet h}}{x_{\bullet\bullet}} = p_{\bullet h}$$

L'écart absolu moyen (mean deviation) est donné par

$$d_g = \frac{\sum_t \left| g_t - \mu_g \right|}{x_{\bullet\bullet}} = \frac{\sum_t \left| g_t - p_{\bullet h} \right|}{x_{\bullet\bullet}} = \frac{\sum_{t \text{ tel que } g_t = 1} \left| g_t - p_{\bullet h} \right| + \sum_{t \text{ tel que } g_t = 0} \left| g_t - p_{\bullet h} \right|}{x_{\bullet\bullet}}$$

$$d_g = \frac{p_{\bullet h} x_{\bullet\bullet} \left| 1 - p_{\bullet h} \right| + \left(1 - p_{\bullet h}\right) x_{\bullet\bullet} \left| 0 - p_{\bullet h} \right|}{x_{\bullet\bullet}} = p_{\bullet h} \left| 1 - p_{\bullet h} \right| + \left(1 - p_{\bullet h}\right) \left| 0 - p_{\bullet h} \right|$$

$$d_g = 2 p_{\bullet h} \left(1 - p_{\bullet h}\right)$$

La variance, quant à elle, est donnée par

$$\sigma_g^2 = \frac{\sum_t \left(g_t - p_{\bullet h}\right)^2}{x_{\bullet\bullet}} = \frac{\sum_t \left(g_t^2 - 2 s_t p_{\bullet h} + p_{\bullet h}^2\right)}{x_{\bullet\bullet}} = \frac{\sum_t g_t^2 - 2 p_{\bullet h} \sum_t g_t + \sum_t p_{\bullet h}^2}{x_{\bullet\bullet}}$$

$$\sigma_g^2 = \frac{\sum_t g_t - 2 p_{\bullet h} \sum_t g_t + \sum_t p_{\bullet h}^2}{x_{\bullet\bullet}} = \frac{p_{\bullet h} x_{\bullet\bullet} - 2 p_{\bullet h} \left(p_{\bullet h} x_{\bullet\bullet}\right) + p_{\bullet h}^2 x_{\bullet\bullet}}{x_{\bullet\bullet}}$$

$$\sigma_g^2 = p_{\bullet h} - p_{\bullet h}^2 = p_{\bullet h} \left(1 - p_{\bullet h}\right)$$

Le coefficient de localisation et l'indice de dissimilarité : pas la même chose !

En science régionale, le coefficient de localisation [14] est largement utilisé pour mesurer le degré de spécificité de la répartition spatiale d'une activité économique par rapport à l'ensemble.

Dans une table de contingence de l'emploi par zone et par branche, $p_{i/\bullet h}$ désigne la fraction de l'emploi total de la branche *h* qui est situé dans la zone *i* ; et $p_{i\bullet}$ désigne la fraction de l'emploi total de l'ensemble des branches qui est situé dans la zone *i*. Le coefficient de localisation se définit comme

$$CL = \frac{1}{2} \sum_i \left| p_{i/\bullet h} - p_{i\bullet} \right|$$

14 Selon Isard (1960, p. 251), c'est à P. Sargant Florence que l'on doit l'introduction du coefficient de localisation parmi les outils de la science régionale ; Duncan et Duncan (1955) citent P. Sargant FLORENCE, W. G. FRITZ et R. C. GILLES, « Measures of industrial distribution », chap. 5 dans : National Resources Planning Board, *Industrial Location and National Resources*, Washington, Government Printing Office, 1943.

À première vue, c'est un indice de dissimilarité. Mais non ! En vérité, la relation entre le coefficient de localisation *CL* et l'indice de dissimilarité *D* est donnée par

$$CL = \left(1 - p_{\bullet h}\right)D$$

Démonstration :

D étant appliqué à une dichotomie, on a

$$D = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}} \right|$$

$$D = \frac{1}{2\left(1 - p_{\bullet h}\right)}\sum_i \left| p_{i/\bullet h}\left(1 - p_{\bullet h}\right) - p_{i\bullet} + p_{ih} \right|$$

$$D = \frac{1}{2\left(1 - p_{\bullet h}\right)}\sum_i \left| p_{i/\bullet h} - p_{i/\bullet h}\, p_{\bullet h} - p_{i\bullet} + p_{ih} \right|$$

$$D = \frac{1}{2\left(1 - p_{\bullet h}\right)}\sum_i \left| p_{i/\bullet h} - p_{ih} - p_{i\bullet} + p_{ih} \right|$$

$$D = \frac{1}{2\left(1 - p_{\bullet h}\right)}\sum_i \left| p_{i/\bullet h} - p_{i\bullet} \right| = \frac{CL}{\left(1 - p_{\bullet h}\right)}$$

La différence vient de ce que le coefficient de localisation compare la distribution d'un groupe (une branche d'activité) avec celle de l'ensemble dont ce groupe fait partie, alors que l'indice de dissimilarité compare la distribution d'un groupe avec celle du reste de la population (les *autres* activités). Il s'ensuit que l'on ne peut pas donner au coefficient de localisation l'interprétation métaphorique que l'on donne à l'indice de dissimilarité, en termes de la fraction du groupe à déplacer pour obtenir des distributions identiques. En outre, le domaine de variation de *CL*, de zéro à $\left(1 - p_{\bullet h}\right)$, est plus étroit pour les branches plus importantes, de sorte qu'il est difficile de comparer les coefficients de localisation de branches de différentes tailles. Par contre, si l'on veut mesurer à quel point la répartition spatiale de chaque activité économique est particulière à cette activité-là, l'indice de dissimilarité présente l'inconvénient d'utiliser une distribution de référence différente pour chaque branche : c'est celle de l'ensemble des autres branches, un ensemble qui est défini différemment pour chaque branche, évidemment.

On peut illustrer ces différences à l'aide de l'exemple utilisé au début de ce chapitre.

Emploi par zone et par branche et distribution de l'emploi entre zones

	Emploi					***Distribution entre zones***				
BRANCHE	*B*1	*B*2	*B*3	*B*1+2	Total	*B*1	*B*2	*B*3	*B*1+2	Total
ZONE										
*Z*1	48	325	287	373	660	0,400	0,542	0,598	0,518	0,550
*Z*2	27	185	148	212	360	0,225	0,308	0,308	0,294	0,300
*Z*3	45	90	45	135	180	0,375	0,150	0,094	0,188	0,150
Total	120	600	480	720	1200	1,000	1,000	1,000	1,000	1,000

Comparaison de la distribution géographique de la branche *B*3
avec celle de l'ensemble des trois branches, puis avec la somme de *B*1 et *B*2

BRANCHE	*B*3	Total	Dif.absol.	*B*1+2	Dif.absol.
ZONe					
*Z*1	0,598	0,550	0.048	0,518	0,080
*Z*2	0,308	0,300	0.008	0,294	0,014
*Z*3	0,094	0,150	0.056	0,188	0,094
Total	1,000	1,000	0,113	1,000	0,188

Appliquons donc la formule de calcul de l'indice de dissimilarité à chacune des deux comparaisons. Dans le premier cas (*B*3 et total), on obtient le coefficient de localisation :

$$CL = \frac{|0{,}048| + |0{,}008| + |-0{,}056|}{2} = 0{,}056$$

Dans le second cas (*B*3 et *B*1+2), on obtient l'indice de dissimilarité :

$$D = \frac{|0{,}080| + |0{,}014| + |-0{,}094|}{2} = 0{,}094$$

Les résultats numériques sont bel et bien différents, comme prévu. Mais ils sont néanmoins liés par la relation

$$CL = \left(1 - \frac{480}{1200}\right)D = 0{,}6 \times 0{,}094 = 0{,}056$$

où le facteur 0,6 est égal à la part de l'emploi des branches *autres que B*3.

Lorsque la sous-population ne représente qu'une petite fraction de la population parente, $p_{\bullet h}$ est petit et la valeur du coefficient de localisation est proche de celle de l'indice de dissimilarité.

Dans le cas particulier où il y a ségrégation totale, l'indice de dissimilarité *D* est égal à 1 et le coefficient de localisation est égal à la part de l'emploi des branches *autres que B*3. On peut illustrer ce dernier point à l'aide de l'exemple de ségrégation totale déjà étudié.

Coefficient de localisation : exemple de ségrégation totale

ETHNIE	Nombres Martiens x_i	Nombres Total y_i	Répartitions Martiens v_i	Répartitions Total w_i	Écart $\lvert v_i - w_i \rvert$
PLANÈTE					
TERRE	0	6	0,00	0,40	0,40
LUNE	0	2	0,00	0,13	0,13
MARS	3	3	0,43	0,20	0,23
JUPITER	4	4	0,57	0,27	0,30
TOTAL	7	15	1,00	1,00	

Coefficient de localisation :

$$\frac{0{,}40 + 0{,}13 + 0{,}23 + 0{,}30}{2} = 0{,}53 = 1 - \frac{7}{15}$$

$=$ fraction de non-Martiens dans la population $=$ fraction de Terriens

On obtiendrait de même pour les Terriens un coefficient de localisation de

$$0{,}47 = 1 - \frac{8}{15}$$

Post scriptum : le coefficient de localisation et les quotients de localisation

À cause de la ressemblance entre leur noms, on peut être porté à confondre le coefficient de localisation et le quotient de localisation. Mais alors que le coefficient de localisation compare deux distributions, le quotient de localisation compare deux parts (voir ci-haut), c'est-à-dire deux points correspondants sur deux distributions. Il y a cependant une relation entre les deux, que l'on trouve en développant la définition du coefficient de localisation :

$$CL = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - p_{i\bullet} \right| = \frac{1}{2}\sum_i p_{i\bullet} \left| \left(\frac{p_{i/\bullet h}}{p_{i\bullet}} \right) - 1 \right| = \frac{1}{2}\sum_i p_{i\bullet} \left| QL_{ih} - 1 \right|$$

Le coefficient de localisation est une moyenne pondérée des écarts absolus entre les quotients de localisation et la valeur repère 1.

UN DERNIER REGARD CRITIQUE

De même que l'on peut construire des indices de prix dont les fondements théoriques sont plus satisfaisants que ceux des indices de Laspeyres et de Paasche, moyennant un degré accru de complication, on peut définir des indicateurs de dissimilarité plus raffinés. Waldorf (1993) en fournit un exemple. Il sied cependant de s'interroger, selon le contexte, sur la pertinence de tels

raffinements et sur leur portée concrète. En outre, la présentation de Waldorf (1993) n'évite pas complètement le piège qui consiste à glisser de la métaphore à l'interprétation littérale : dans le contexte d'une étude de la ségrégation raciale aux États-Unis, l'auteur évoque une mesure de l'« effort requis » par un déplacement de la population.

1-5.3 Distance et dissimilarité

Parmi les mesures de dissimilarité, certaines sont des mesures de distances, généralisées à plus de deux ou trois dimensions, en ce sens qu'elles possèdent les propriétés que doit avoir une mesure de distance. Réciproquement, on peut considérer la distance comme un cas particulier de la dissimilarité : la distance est une dissimilarité entre deux objets par rapport à leur situation dans l'espace ou plus simplement entre deux lieux dans l'espace.

Une surface (comme la surface de la terre, si l'on ignore le relief [15]) est un espace à deux dimensions. La spécification d'une situation dans l'espace comporte donc deux dimensions : longitude et latitude, ou coordonnées cartésiennes (x,y). Par conséquent, la mesure de la distance géographique comprend elle aussi deux dimensions. Et, même si, dans la vie courante, on utilise sans y penser la distance euclidienne, il y a plus d'une façon de mesurer la distance [16].

Une mesure de distance doit satisfaire certaines conditions. La fonction $d(a,b)$ est une fonction de distance si et seulement si, pour tout ensemble de lieux a, b et c, elle satisfait les quatre conditions suivantes :

(c1) non négativité :

$$d(a,b) \geq 0$$

(c2) identité :

$d(a,b) = 0$ si, et seulement si, $a = b$

(c3) symétrie :

$$d(a,b) = d(b,a)$$

(c4) inégalité triangulaire :

$$d(a,c) \leq d(a,b) + d(b,c)$$

---

15 Si l'on tient compte du relief, on a un espace tri-dimensionnel.

16 Voir Huriot et Perreur (1990 et 1994).

La mesure de distance la plus familière est la *distance euclidienne*. La distance euclidienne entre le point *a*, de coordonnées (x_a,y_a), et le point *b*, de coordonnées (x_b,y_b), est donnée par :

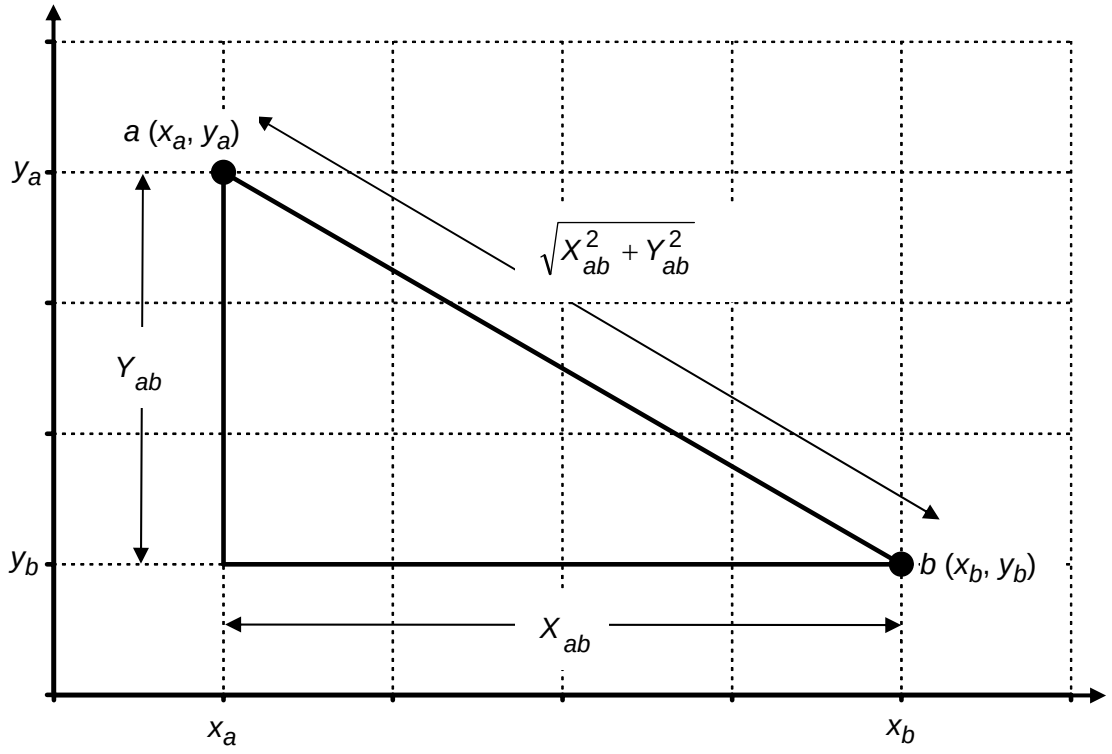
$$d_e(a,b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

Parmi les autres mesures de distance, signalons la *distance rectilinéaire* [aussi appelée distance *rectangulaire* ou distance (*selon la métrique*) *de Manhattan* ; voir Huriot et Perreur, 1994, p. 44] :

$$d_r(a,b) = \left| x_a - x_b \right| + \left| y_a - y_b \right|$$

La distance selon la métrique de Manhattan est la distance qu'il faut parcourir pour aller de *a* à *b* en suivant le tracé des rues, lorsque celles-ci forment une grille rectangulaire comme à Manhattan.

Les deux métriques sont illustrées dans la figure qui suit, où $X_{ab} = |x_a - x_b|$ et $Y_{ab} = |y_a - y_b|$



Puisque la distance géographique peut être interprétée comme une dissimilarité, la réciproque est aussi vraie : les mesures de la distance peuvent être utilisées pour mesurer des dissimilarités qui ne sont pas des distances géographiques.

Ainsi, considérons deux objets que l'on décrit à l'aide de n variables, dont chacune mesure une caractéristique (dimension) pertinente :

$x_{11}, x_{12}, ..., x_{1n}$ pour le premier objet et

$x_{21}, x_{22}, ..., x_{2n}$ pour le second.

Exemple :

> Si les deux objets étaient deux quartiers d'une ville, les caractéristiques pertinentes pourraient être la densité de la population, la proportion de la population ayant moins de quinze ans, la proportion de la population ayant complété l'école primaire, le revenu moyen des ménages, etc.

Pour mesurer la dissimilarité entre deux objets multidimensionnels, on utilise souvent la distance euclidienne généralisée, qui est définie par :

$$\sqrt{\sum_i \left(x_{1i} - x_{2i}\right)^2}$$

On utilise aussi la distance rectilinéaire généralisée ou distance généralisée selon la métrique de Manhattan, qui est donnée par

$$\sum_i \left| x_{1i} - x_{2i} \right|$$

Le lecteur perspicace aura remarqué la parenté entre l'indice de dissimilarité D et la distance généralisée selon la métrique de Manhattan : D est égal à la moitié de la distance rectilinéaire généralisée. Dans le présent contexte cependant, les deux objets comparés ne sont pas nécessairement des distributions. Il s'ensuit notamment qu'il n'y a pas de valeur maximum inhérente à la distance rectilinéaire généralisée (ni à la distance euclidienne généralisée, d'ailleurs).

De façon générale, la valeur d'une mesure de distance n'est pas indépendante des unités de mesure des variables sous-jacentes. C'est pourquoi, lorsqu'on compare des objets multidimensionnels à l'aide d'une distance généralisée, on doit affronter un problème analogue à celui auquel on fait face dans la construction d'un nombre indice. En effet, le choix de l'unité de mesure de chaque variable détermine implicitement quel sera son poids dans la mesure de

distance-dissimilarité. C'est seulement quand les objets comparés sont des distributions que le problème des échelles de mesure ne se pose pas.

1-5.4 La mesure de la similarité en statistique

Le problème de la mesure de la similarité se pose souvent en statistique. Par exemple, considérons deux séries d'observations sur deux variables :

$x_1, x_2, ..., x_n$ et $y_1, y_2, ..., y_n$

Le coefficient de corrélation simple est une mesure de la similarité entre ces deux séries de données [17].

De même, pour évaluer l'exactitude d'un modèle par rapport aux données qui ont servi à estimer ses paramètres, on mesure la similarité entre les valeurs observées et les valeurs prédites par le modèle. L'une des mesures les plus utilisées pour cela est le coefficient de détermination multiple R^2 (dont il sera question dans la troisième partie de cet ouvrage).

Enfin, le Khi-deux de Pearson [18] est une mesure de la dissimilarité entre les effectifs observés et les effectifs « théoriques » prédits par une hypothèse.

Toutes ces mesures appartiennent à la grande famille des mesures de similarité et de dissimilarité.

On trouve dans Webber (1984, p. 41-45) une discussion intéressante de la pertinence de différentes mesures d'ajustement (dans le contexte de l'évaluation de l'exactitude du modèle de répartition spatiale de Lowry).

1-5.5 Autres mesures de similarité et de dissimilarité

Il existe une grande abondance de mesures de similarité et de dissimilarité. Legendre et Legendre (1984, tome 2, chap. 6 et 1998, chap. 7) présentent et discutent une multitude de mesures, utilisées en écologie numérique et qui pourraient être employées pour l'analyse spatiale en sciences sociales.

---

17 Voir l'annexe 2-A « Rappel de quelques formules courantes en statistique ». Le coefficient de corrélation mesure plus exactement la similarité entre les valeurs observées d'une variable et ses valeurs prédites à l'aide de l'autre variable.

18 Voir 4-1. Le Khi-deux n'est cependant pas une mesure symétrique : sa valeur change si l'on intervertit les rôles des valeurs observées et des valeurs théoriques.