PROBLÉMATIQUE DE LA MESURE DE LA SIMILARITÉ/DISSIMILARITÉ

Mesure

Mesurer, c'est comparer

Une *mesure* est une correspondance qui permet de comparer deux objets par rapport à une propriété donnée.

Le problème de la multidimensionnalité : les nombres indices

- Lazarsfeld : concept → dimensions → indicateurs (mesures)
- Problème :
  objet ou concept multidimensionnel
  mais on veut le traiter comme un tout
  ⇒ il faut combiner les mesures partielles en une seule mesure globale, qui les résume
- Un nombre indice est une mesure : permet de comparer par rapport au concept
  - Fiabilité ?
  - Validité ?

Comparer sans indice : mesure de la similarité/dissimilarité

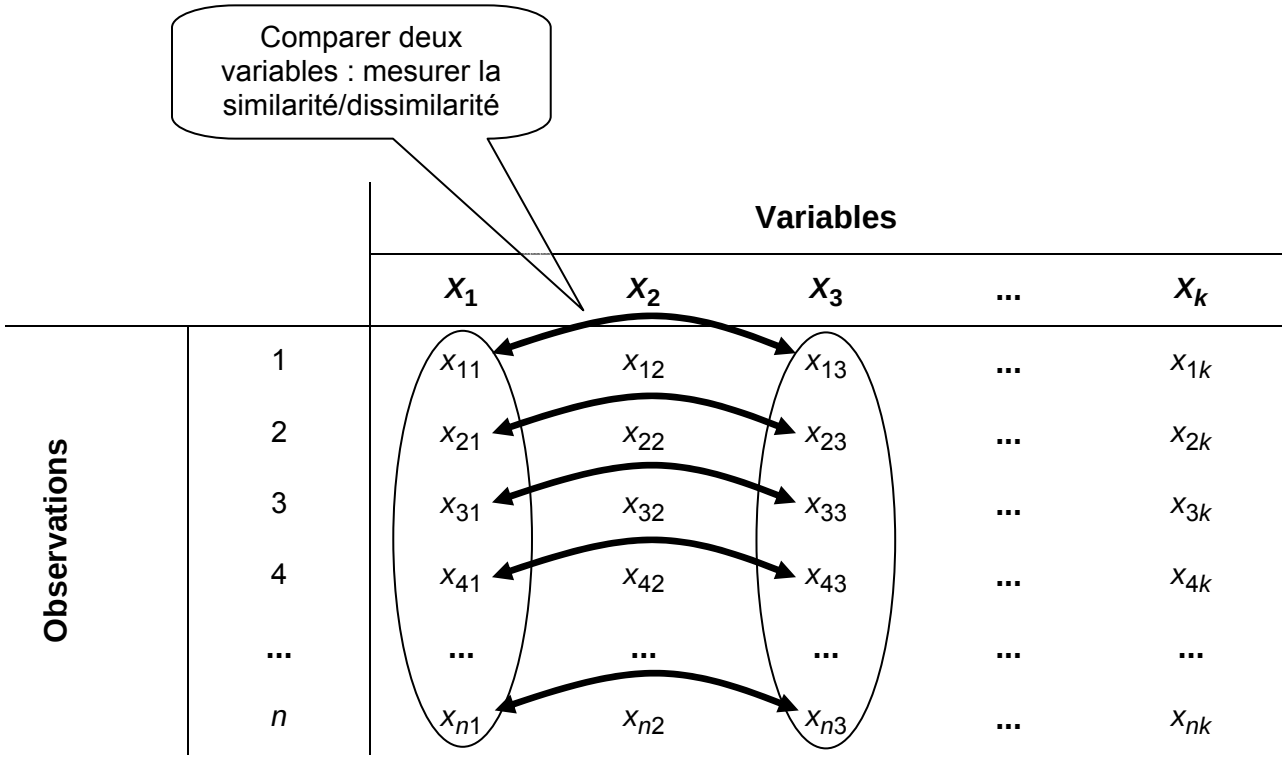
- Certains concepts ne sont pas réductibles à un indice

  On ne peut pas associer une mesure unique au concept
- Dimensions multiples → indicateurs multiples → comparaisons multiples
- Comment faire la synthèse des comparaisons ?

  Au moyen d'un « indice des comparaisons »
- On mesure
  - **non pas** le degré auquel chaque objet « possède » le concept
  - **mais plutôt** le degré de similarité entre les objets par rapport au concept

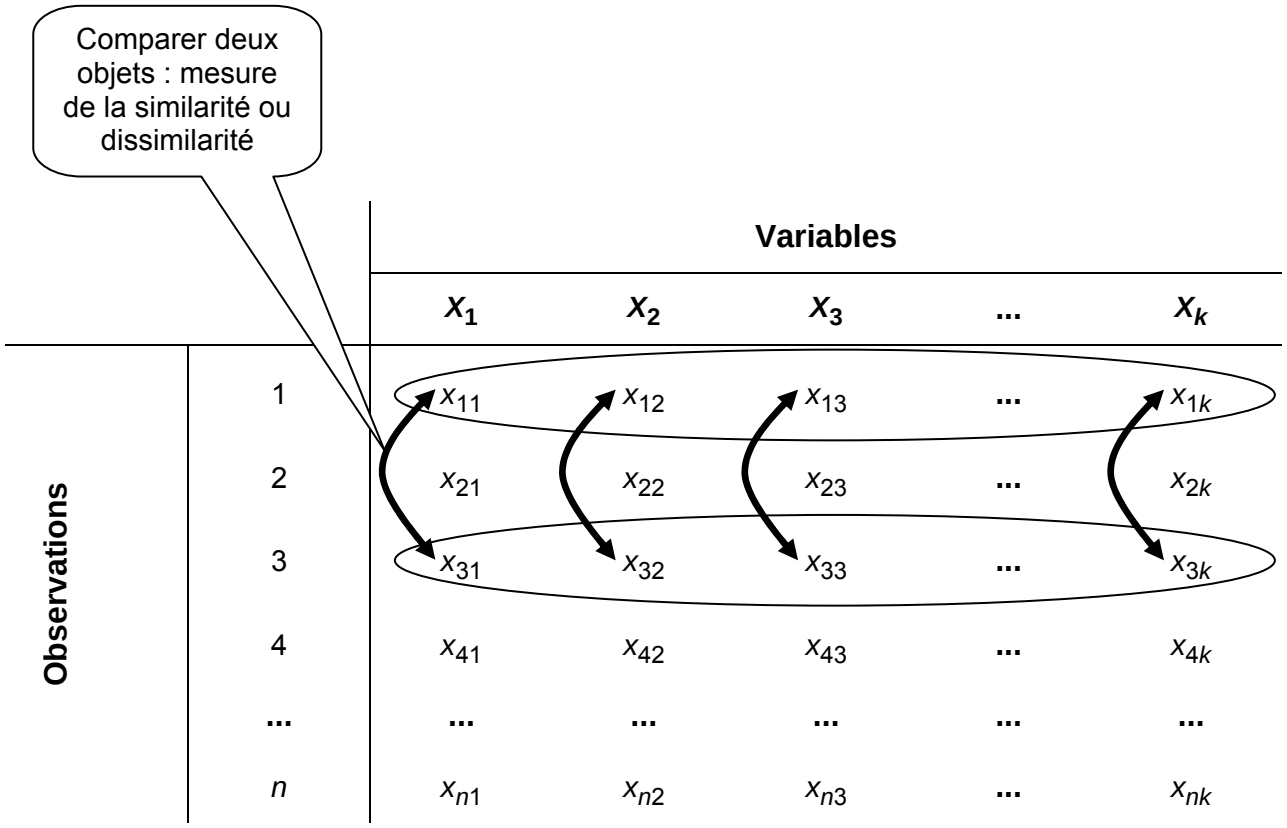
STRUCTURE DES DONNÉES (3)

POINT DE VUE HORIZONTAL : SIMILARITÉ/DISSIMILARITÉ



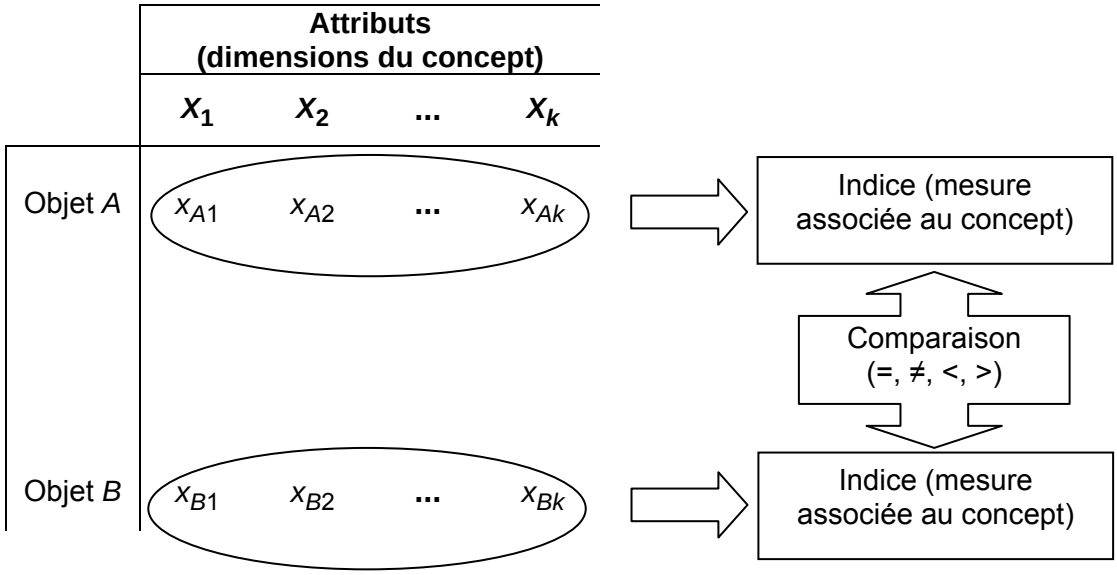
STRUCTURE DES DONNÉES (7)

POINT DE VUE VERTICAL : SIMILARITÉ/DISSIMILARITÉ (BIS)

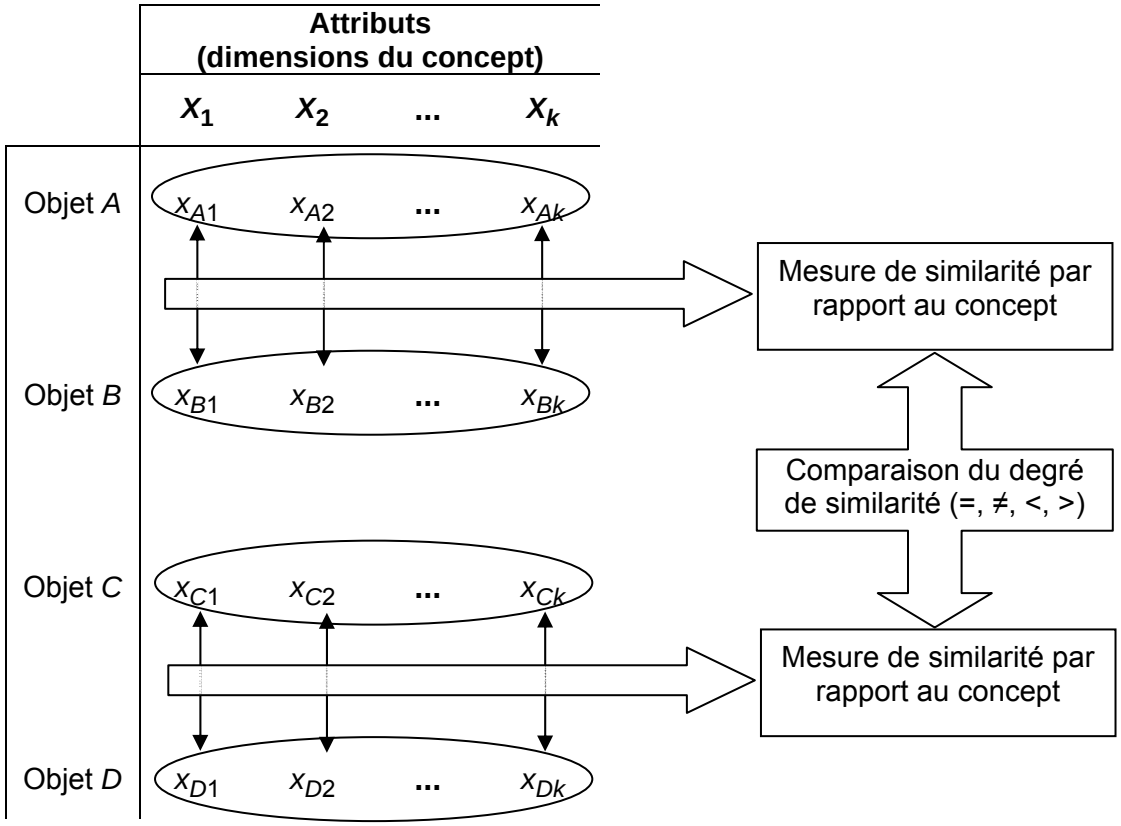


DIFFÉRENCE ENTRE LA MESURE DE LA SIMILARITÉ ET LA CONSTRUCTION D'UN NOMBRE INDICE

Construction d'un nombre indice



Mesure de la similarité



DANS QUELLES CIRCONSTANCES... ?

Mesure de la similarité/dissimilarité en général

- Construction de typologies, algorithmes de classification
- Mesures d'ajustement (« goodnes of fit ») en statistique : similarité entre les observations et les prédictions d'un modèle
  (ex. : fréquences observées et théoriques d'un tableau de contingence;
  coefficient de détermination multiple de la régression)

Mesure de la similarité/dissimilarité dans un tableau de contingence (exemple)

- entre lignes ou entre colonnes...
- quant à la **structure**, c'est-à-dire quant à la répartition

L'indicateur de spécificité (quotient de localisation) s'applique à chaque cellule séparément.

L'analyse des tableaux de contingence, le test d'indépendance et les mesures d'intensité s'appliquent à l'ensemble du tableau.

La mesure de similarité/dissimilarité s'applique à chaque paire de lignes ou de colonnes.

Dans un tableau de contingence : cas particulier...

Plus généralement : Mesure de la similarité/dissimilarité entre deux distributions

- Distribution de fréquences ou distribution d'une variable continue
- En particulier, distributions spatiales
- Dans une distribution, la somme des parts est égale à 1 (100 %) : $\sum_{1} p_i = 1$

  Cela règle le problème de la pondération

Distribution observée et distribution théorique

- Approche analogue à la construction du test d'hypothèse d'indépendance
- **mais** ici, la distribution théorique n'est pas une hypothèse à tester,
  c'est la représentation du degré maximum d'une propriété
- Cette approche s'applique notamment à la...

Mesure de l'inégalité ou de la concentration

- La concentration est le contraire de l'égalité dans une distribution
- Elle peut se mesurer par le degré de dissimilarité par rapport à une distribution de référence qui représente l'égalité parfaite
- Nombreux champs d'application de la mesure de l'inégalité :
  - géographie : concentration spatiale des phénomènes
  - économie : inégalités de revenu et questions de justice sociale; concentration de marché

POPULATION ACTIVE EMPLOYÉE DANS LA RÉGION MÉTROPOLITAINE DE MONTRÉAL ZONE DE RÉSIDENCE, SELON LE SEXE ET LA PROFESSION, 1991

	Professions					
Zone de résidence	Directeurs, gérants, administra-teurs et assimilés	Profession-nels, enseignants et cols blancs spécialisés	Employés de bureau et travailleurs dans la vente	Ouvriers	Travailleurs spécialisés dans les services, personnel d'exploitation des transports, etc.	**TOTAL toutes professions**
Femmes						
Ville de Montréal	24 025	58 204	76 450	24 385	28 825	**211 889**
Reste de la CUM	22 575	42 207	70 003	14 065	17 435	**166 285**
Couronne Nord	16 785	31 699	63 491	11 975	18 630	**142 580**
Couronne Sud	18 365	35 674	65 290	10 485	19 380	**149 194**
Hors RMR	3 265	7 535	11 089	3 190	3 565	**28 644**
Total Femmes	**85 015**	**175 319**	**286 323**	**64 100**	**87 835**	**698 592**
Hommes						
Ville de Montréal	32 336	55 045	43 546	65 340	46 850	**243 117**
Reste de la CUM	39 146	39 920	37 819	46 173	28 749	**191 807**
Couronne Nord	33 287	27 560	31 170	62 852	29 329	**184 198**
Couronne Sud	36 006	32 464	30 600	58 778	29 721	**187 569**
Hors RMR	8 270	8 590	8 270	22 305	9 099	**56 534**
Total Hommes	**149 045**	**163 579**	**151 405**	**255 448**	**143 748**	**863 225**
Total hommes et femmes						
Ville de Montréal	56 361	113 249	119 996	89 725	75 675	**455 006**
Reste de la CUM	61 721	82 127	107 822	60 238	46 184	**358 092**
Couronne Nord	50 072	59 259	94 661	74 827	47 959	**326 778**
Couronne Sud	54 371	68 138	95 890	69 263	49 101	**336 763**
Hors RMR	11 535	16 125	19 359	25 495	12 664	**85 178**
Total H + F	**234 060**	**338 898**	**437 728**	**319 548**	**231 583**	**1 561 817**

Source : Statistique Canada, Recensement de 1991

STRUCTURE DES DONNÉES (6)

POINT DE VUE VERTICAL : INÉGALITÉ, DISTRIBUTION

- Caractériser la distribution
- Mesurer l'inégalité ou la concentration

S'il existe un ordre naturel des observations :

- Analyser des séries temporelles
- Analyser l'autocorrélation temporelle ou spatiale

		Variables				
		$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	**...**	$\boldsymbol{x_k}$
Observations	1	x_{11}	x_{12}	x_{13}	**...**	x_{1k}
	2	x_{21}	x_{22}	x_{23}	**...**	x_{2k}
	3	x_{31}	x_{32}	x_{33}	**...**	x_{3k}
	4	x_{41}	x_{42}	x_{43}	**...**	x_{4k}
	...	**...**	**...**	**...**	**...**	**...**
	n	x_{n1}	x_{n2}	x_{n3}	**...**	x_{nk}

COMMENT MESURER L'INÉGALITÉ ?

Qu'est-ce que l'inégalité ? Exemple du revenu

- Entre 2 personnes :
  - Si $R_1 = R_2 \Rightarrow$ égalité
  - Si $R_1 \neq R_2 \Rightarrow$ inégalité : elle peut se mesurer par

    la différence $(R_1 - R_2)$, le rapport $\frac{R_1}{R_2}$ ou une transformation de ceux-ci.
- Entre plus de 2 personnes :
  - Si $R_1 = R_2 = R_3 = ... \Rightarrow$ égalité
  - Sinon, il n'y a pas égalité **mais** comment mesurer le degré d'inégalité ?

Propriétés désirables d'une mesure d'inégalité (Valeyre, 1993)

1. Non négative
2. Égale à zéro si, et seulement si

   la distribution observée identique à distribution de référence.
3. Toutes observations traitées de la même manière.
4. Indépendante de la valeur moyenne de la variable.

   Indépendante de la taille de la population.
5. L'agrégation d'observations affichant le même degré de spécificité ne doit pas changer la valeur de la mesure.
6. Principe de transfert de Pigou-Dalton : une mesure d'inégalité doit diminuer si la distribution est modifiée d'une façon qui réduit incontestablement l'inégalité.

COMMENT MESURER L'INÉGALITÉ ? (SUITE)

La dispersion des valeurs observées s'interprète souvent comme le reflet de la concentration ou de l'inégalité de la propriété mesurée.

- Exemple : avec des données sur le revenu, si tout le monde a le même revenu, il n'y a pas de dispersion (la variance est nulle), il n'y a pas d'inégalité entre les individus (observations) et le revenu n'est pas concentré ; plus il y a de différences entre les revenus, plus la variance est grande.

Rappel : mesures de dispersion en statistique descriptive

- Domaine de variation : valeur minimum et valeur maximum
- Écart inter-quartile
- Variance : $\sigma_X^2 = \frac{1}{n}\sum_i (x_i - \mu_X)^2$

  NOTE : Cette formule est celle qui s'applique à une population, puisque la statistique descriptive ne distingue pas entre population et échantillon.
- Écart-type : $\sigma_X = \sqrt{\sigma_X^2}$
- Coefficient de variation : $C_X = \frac{\sigma_X}{\mu_X}$

Seul le coefficient de variation possède les 6 propriétés désirées.

Peut-on mesurer l'inégalité ou la concentration sans référer à la moyenne ?

- Oui ! Corrado Gini (1884-1965) a proposé de comparer chacun des individus avec chacun des autres : cela donne la différence moyenne de Gini.

Autres mesures d'inégalité ou de concentration

- Le coefficient de concentration de l'économie industrielle

  $C4 = \sum_{i=1}^{4} s_i$, où s_i est la part de *i* dans le total
- L'indice de concentration de Hirschman-Herfindahl

  $H = \sum_{i=1}^{n} s_i^2$

  $\frac{1}{n} \le H \le 1$

  Interprétation en «nombre équivalent»

  Variance des parts = $\frac{1}{n}\sum_{i=1}^{n}\left(s_i - \frac{1}{n}\right)^2 = \frac{H}{n} - \frac{1}{n^2}$
- Mesure d'entropie

L'INDICE DE CONCENTRATION DE GINI :

LA DIFFÉRENCE MOYENNE DE GINI (EXEMPLE NUMÉRIQUE)

Données (fictives)

Individus	Revenu
A	100
B	40
C	30
D	20
E	20
F	20
G	20
H	20
I	20
J	10

Total	300
Moyenne	30
Éc. type	24,49
Coef. var.	0,816

Calcul de la différence (absolue) moyenne $\left|x_i - x_j\right|$

	A	***B***	***C***	***D***	***E***	***F***	***G***	***H***	***I***	***J***
	100	***40***	***30***	***20***	***20***	***20***	***20***	***20***	***20***	***10***
100	0	60	70	80	80	80	80	80	80	90
40	60	0	10	20	20	20	20	20	20	30
30	70	10	0	10	10	10	10	10	10	20
20	80	20	10	0	0	0	0	0	0	10
20	80	20	10	0	0	0	0	0	0	10
20	80	20	10	0	0	0	0	0	0	10
20	80	20	10	0	0	0	0	0	0	10
20	80	20	10	0	0	0	0	0	0	10
20	80	20	10	0	0	0	0	0	0	10
10	90	30	20	10	10	10	10	10	10	0

Somme	2000
Dif. Moy. Gini	20
Coef. Gini	0,333

$$\Delta = \frac{1}{N^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|, \text{ avec } N = 10 : \Delta = \frac{2000}{10^2} = 20$$

$$G = \frac{\Delta}{2\mu} = \frac{20}{2 \times 30} = 0{,}333$$

L'INDICE DE CONCENTRATION DE GINI :
LA DIFFÉRENCE MOYENNE DE GINI
(EXEMPLE NUMÉRIQUE AVEC DONNÉES GROUPÉES)

Données groupées

		f_i	y_i
Cat.	**Rev.**	**N.**	**R/N**
>25	170	3	56,67
15-25	120	6	20
<15	10	1	10
Tot.	300	10	
Moy.	30	(pondérée)	

Écarts $\left| y_i - y_j \right|$

		>25	**15-25**	**<15**
		56,67	***20***	***10***
>25	***56,67***	0	36,67	46,67
15-25	***20***	36,67	0	10
<15	***10***	46,67	10	0

Poids $f_i f_j$

		>25	**15-25**	**<15**
		3	***6***	***1***
>25	***3***	9	18	3
15-25	***6***	18	36	6
<15	***1***	3	6	1
Tot.		**100**		

Écarts pondérés $\left| y_i - y_j \right| f_i f_j$

	>25	**15-25**	**<15**
>25	0	660	140
15-25	660	0	60
<15	140	60	0
Tot.	**1720**		

$$\Delta = \frac{1}{N^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| y_i - y_j \right| f_i f_j = \frac{1720}{100} = 17{,}2$$

$$G = \frac{\Delta}{2\mu} = \frac{17{,}2}{2 \times 30} = 0{,}287$$

L'INDICE DE CONCENTRATION DE GINI : LA DIFFÉRENCE MOYENNE DE GINI (FORMULE ALGÉBRIQUE)

Notation

n = le nombre de valeurs distinctes observées

f_j = fréquence de la valeur y_j dans la distribution

$N = \sum_{j=1}^{n} f_j$ = nombre d'observations

Définition de la différence moyenne de Gini

$$\Delta = \frac{1}{N^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \left| y_j - y_k \right| f_j \, f_k$$

Observations groupées par classes

y_j = valeur *moyenne* de la variable Y dans la classe j

$v_j = \frac{f_j}{N}$, la fraction de la population appartenant à la classe j.

La valeur moyenne de la variable Y s'écrit alors

$$\mu = \frac{1}{N} \sum_{j=1}^{n} f_j \, y_j = \sum_{j=1}^{n} v_j y_j$$

Notation

$M = \mu N = \sum_{j=1}^{n} f_j y_j$ = somme des valeurs de la variable Y

$w_j = \frac{f_j y_j}{\sum_{k=1}^{n} f_k y_k} = \frac{f_j y_j}{N\mu} = \frac{v_j y_j}{\mu}$ = fraction de la somme allouée à la classe j.

$Cw_j = \sum_{k=1}^{j} w_j$, avec observations par ordre croissant des w_j / v_j

$$\Delta = 2\mu \left(1 - \sum_{j=1}^{n} v_j \, Cw_j \; - \sum_{j=1}^{n} v_j \, Cw_{j-1} \right)$$

CALCUL DE L'INDICE DE CONCENTRATION DE GINI

$v_j = \dfrac{f_j}{N}$, la fraction de la population appartenant à la classe *j*.

$Cw_j = \sum_{k=1}^{j} w_j$, avec observations par ordre croissant des w_j / v_j

$$G = \frac{\Delta}{2\mu} = 1 - \left(\sum_{j=1}^{n} v_j \, Cw_j \;\; + \sum_{j=1}^{n} v_j \, Cw_{j-1} \right) = 1 - \sum_{j=1}^{n} v_j \left(Cw_j + Cw_{j-1} \right)$$

Calcul équivalent d'Arriaga (1975, p. 65-71)

$$G = \sum_{i=2}^{n} Cw_i \, Cv_{i-1} - \sum_{i=2}^{n} Cw_{i-1} \, Cv_i$$

où $Cv_j = \sum_{k=1}^{j} v_k$

LA COURBE DE LORENZ

La courbe de Lorenz

Notation supplémentaire

$Cv_j = \sum_{k=1}^{j} v_k =$ fraction cumulée de la population X

$Cw_j = \sum_{k=1}^{j} w_k =$ fraction cumulée de la population Y

$Cv_n = Cw_n = 1$

Méthode de construction de la courbe de Lorenz

1. Calculer les rapports $\frac{w_i}{v_i}$. Ce sont les *spécificités* associées aux observations.
2. Réordonner les catégories en ordre croissant de $\frac{w_i}{v_i}$: $\frac{w_1}{v_1} < \frac{w_2}{v_2} < \cdots < \frac{w_n}{v_n}$
3. La courbe de Lorenz est l'ensemble des points
   (Cv_i, Cw_i), où les Cv_i sont repérés sur l'axe horizontal.

Propriétés de la courbe de Lorenz

1. $Cv_0 = Cw_0 = 0$
2. $Cv_n = Cw_n = 1$
3. Lorsque les deux distributions sont identiques, on a, pour tout i,
   $Cv_i = Cw_i$
   La courbe de Lorenz coïncide avec la diagonale.
4. $Cv_i \geq Cw_i$ pour i différent de 0 et de n
5. La pente de chaque segment de la courbe de Lorenz est égale à la valeur l'indicateur de spécificité associé à l'observation correspondante :
   pente du segment $i = \frac{Cw_i - Cw_{i-1}}{Cv_i - Cv_{i-1}} = \frac{w_i}{v_i}$
6. La courbe de Lorenz est concave vers le haut, c'est-à-dire que chaque segment a une pente plus abrupte que le précédent : cela découle de 5, puisque, par construction, $\frac{w_i}{v_i} < \frac{w_{i+1}}{v_{i+1}}$

CONSTRUCTION D'UNE COURBE DE LORENZ (EXEMPLE NUMÉRIQUE TIRÉ DE TAYLOR, 1977, P. 180)

Première étape : calcul des w_i / v_i

Zone	x_i Nombre de ménages de classe moyenne	v_i Distrib. de x	y_i Nombre de votes du parti Républ.	w_i Distrib. de y	w_i/v_i
A	30	0,25	30	0,30	1,20
B	20	0,17	15	0,15	0,90
C	10	0,08	8	0,08	0,96
D	10	0,08	5	0,05	0,60
E	20	0,17	19	0,19	1,14
F	30	0,25	23	0,23	0,92
Tot.	**120**	**1,00**	**100**	**1,00**	

Deuxième étape : tri par ordre croissant des w_i/v_i

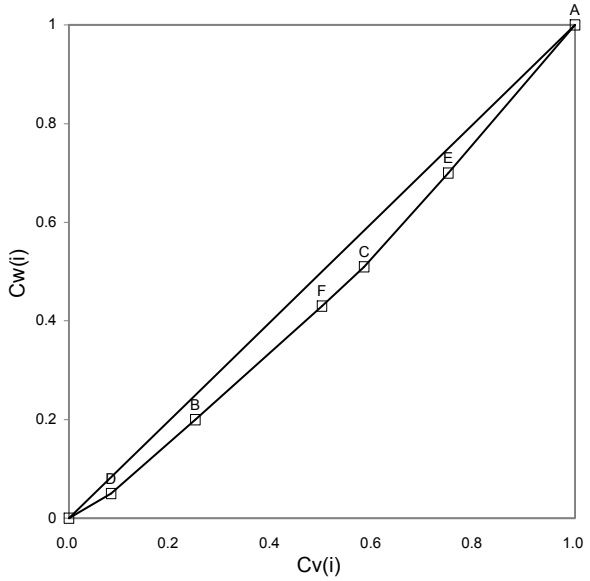
Troisième étape : calcul des Cv_i (abscisses) et des Cw_i (ordonnées)

Zone	x_i	v_i	y_i	w_i	w_i/v_i	Cv_i Abscisse	Cw_i Ordonnée	Différ. $(Cv_i - Cw_i)$	Différ.abs. $\lvert v_i - w_i \rvert$
						0,00	0,00		
D	10	0,08	5	0,05	0,60	0,08	0,05	0,033	0,033
B	20	0,17	15	0,15	0,90	0,25	0,20	0,050	0,017
F	30	0,25	23	0,23	0,92	0,50	0,43	0,070	0,020
C	10	0,08	8	0,08	0,96	0,58	0,51	***0,073***	0,003
E	20	0,17	19	0,19	1,14	0,75	0,70	0,050	0,023
A	30	0,25	30	0,30	1,20	1,00	1,00	0,000	0,050
Tot.	**120**	**1,00**	**100**	**1,00**					**0,147**

Note : on peut voir que le maximum de la différence absolue entre la courbe de Lorenz et la diagonale est égal à $\frac{1}{2}\sum_i \left| v_i - w_i \right|$.

Source : Lorenz.wb3

Courbe de Lorenz



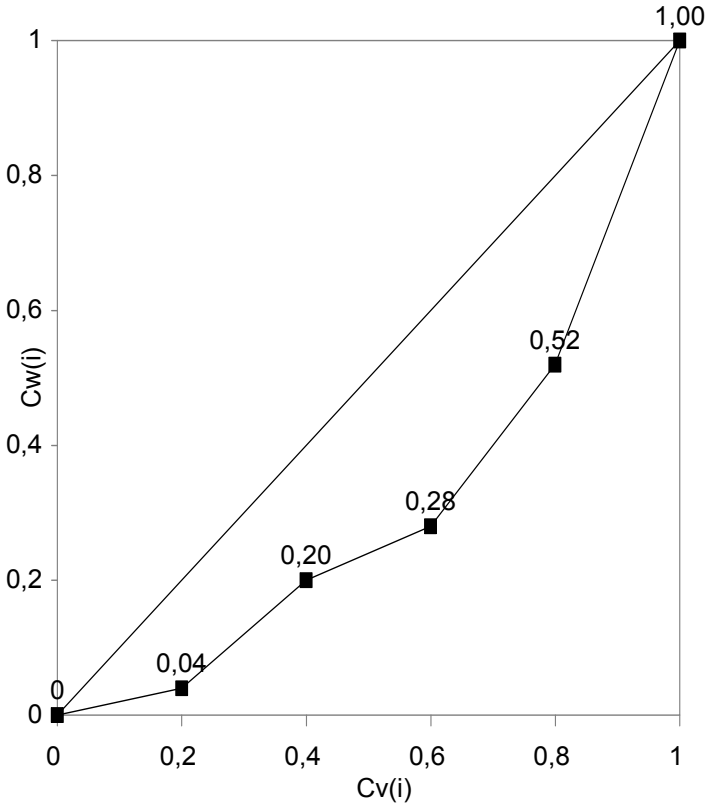
Quatrième étape : calcul de l'indice de concentration de Gini

Zone	x_i	v_i	y_i	w_i	w_i/v_i	Cv_i	Cw_i	$v_i\,Cw_i$	$v_i\,Cw_{i-1}$
						Abscisse	Ordonnée		
						0,00	0,00		
D	10	0,08	5	0,05	0,60	0,08	0,05	0,004	0,000
B	20	0,17	15	0,15	0,90	0,25	0,20	0,033	0,008
F	30	0,25	23	0,23	0,92	0,50	0,43	0,108	0,050
C	10	0,08	8	0,08	0,96	0,58	0,51	0,043	0,036
E	20	0,17	19	0,19	1,14	0,75	0,70	0,117	0,085
A	30	0,25	30	0,30	1,20	1,00	1,00	0,250	0,175
Tot.	**120**	**1,00**	**100**	**1,00**				**0,554**	**0,354**

$$G = 1 - (0{,}554 + 0{,}354) = 0{,}092$$

Source : Lorenz.wb3

LA COURBE DE LORENZ ET LE COEFFICIENT GINI :
CECI N'EST PAS UNE COURBE DE LORENZ !



Source : Lorenz2.wb3, Lorenz2

CALCUL GÉOMÉTRIQUE DE L'INDICE DE CONCENTRATION DE GINI

Définition géométrique de l'indice de concentration de Gini

$$G = \frac{\text{Superficie comprise entre la diagonale et la courbe de Lorenz}}{\text{Superficie totale sous la diagonale}}$$

Calcul

Superficie totale du triangle sous la diagonale = $\frac{Cw_n \times Cv_n}{2} = \frac{1 \times 1}{2} = \frac{1}{2}$

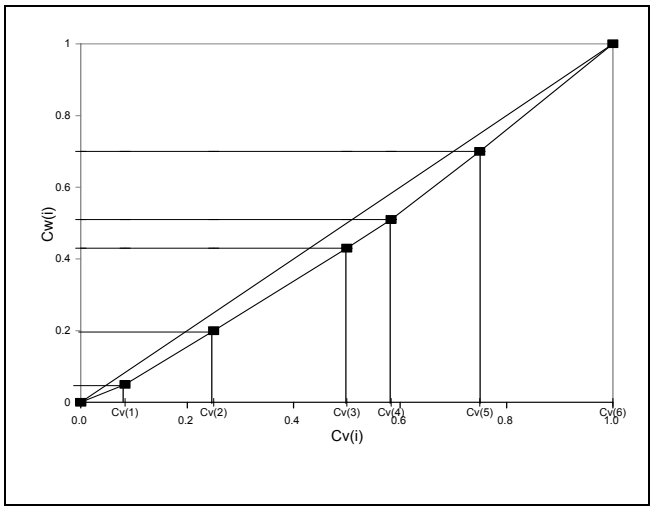
Superficie entre diagonale et courbe de Lorenz = différence entre :

Superficie totale du triangle sous la diagonale (=1/2) et

Superficie sous la courbe de Lorenz

Superficie sous la courbe de Lorenz = somme de n trapèzes :

$$\frac{1}{2} v_i \left(Cw_i + Cw_{i-1}\right)$$

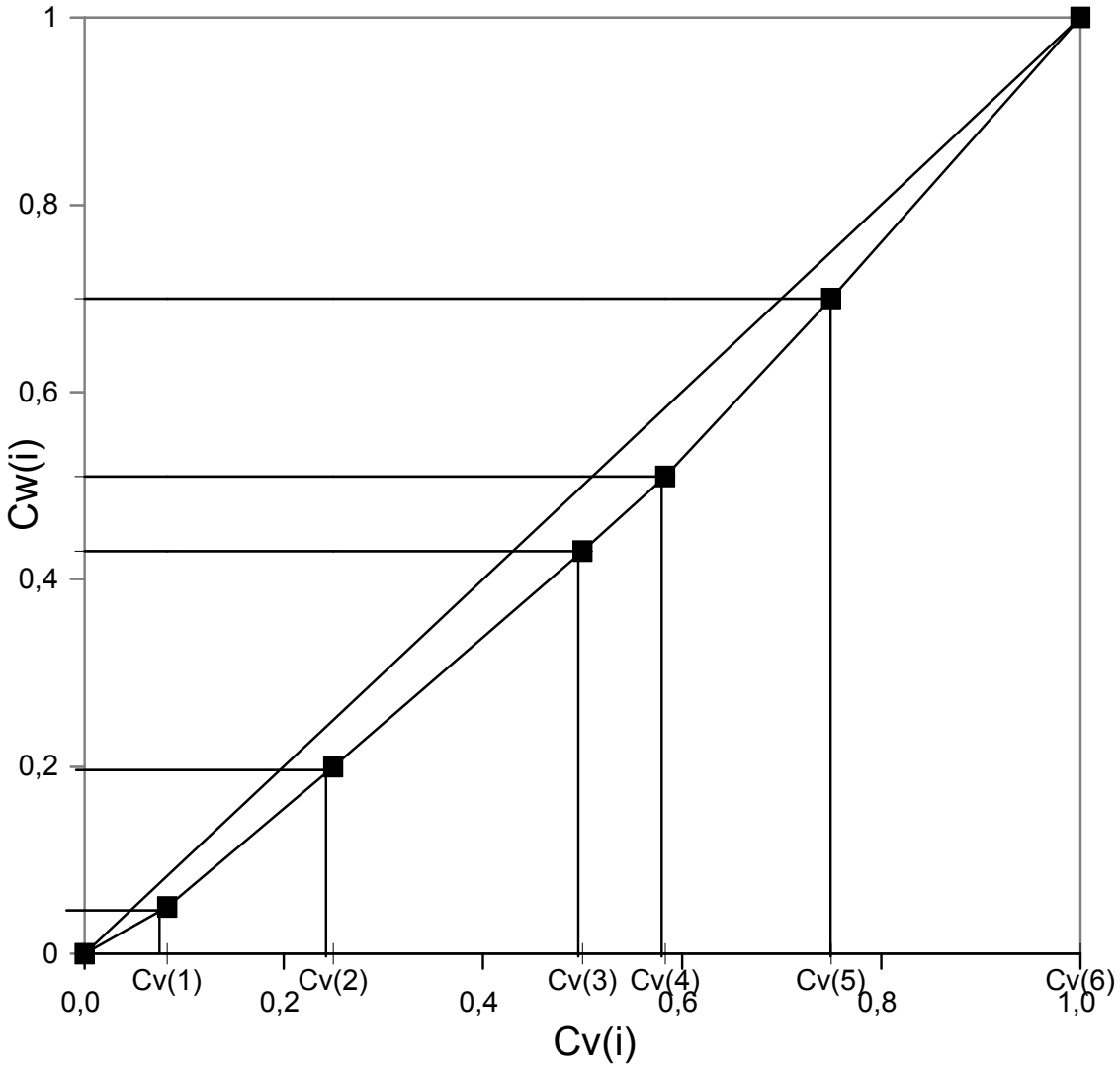


Coefficient Gini :

$$G = \frac{\left(\frac{1}{2}\right) - \left(\frac{1}{2}\sum_{i=1}^{n} v_i \left(Cw_i + Cw_{i-1}\right)\right)}{\left(\frac{1}{2}\right)} = 1 - \sum_{i=1}^{n} v_i \left(Cw_i + Cw_{i-1}\right) = \frac{\Delta}{2\mu}$$

Source : Lorenz.wb3, CalGini2

LA COURBE DE LORENZ ET L'INDICE DE CONCENTRATION DE GINI : CALCUL GÉOMÉTRIQUE



Source : Lorenz.wb3, CalGini2

INTERPRÉTATION ET PROPRIÉTÉS DE L'INDICE DE CONCENTRATION DE GINI

Interprétation

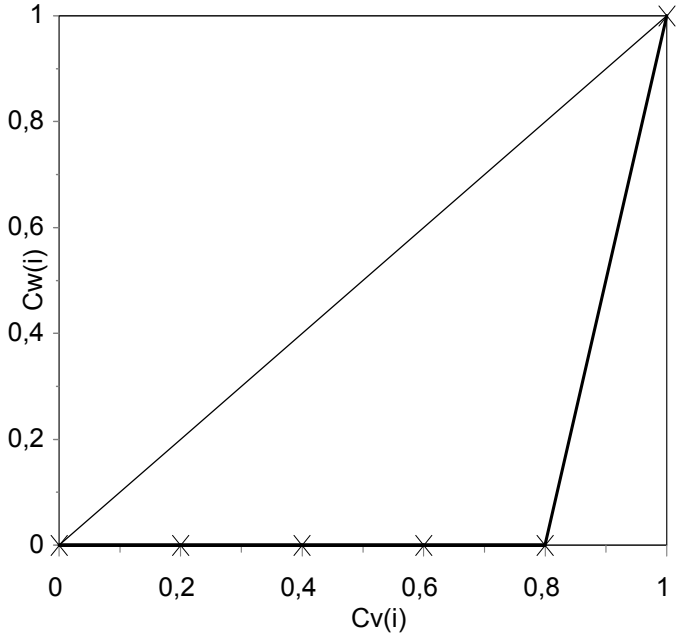
1. Mesure de dissimilarité entre deux distributions
2. Mesure de concentration :
   - V = distribution de référence (axe horizontal)
   - W = distribution dont on veut mesurer la concentration (axe vertical)

Propriétés de l'indice de concentration de Gini

1. Possède les 6 propriétés désirables d'une mesure d'inégalité (Valeyre, 1993)
2. $0 \leq G \leq 1$, ou plus exactement $0 \leq G \leq 1 - v_n$
3. G est symétrique.
4. Quand les données sont regroupées, G est sensible à la définition et au nombre des catégories utilisées (classes, zones).

   Cela se manifeste notamment par : l'agrégation de deux ou plusieurs catégories entraîne une diminution de la valeur de l'indice de Gini, **sauf** si les catégories ont la même spécificité.
5. En tant que mesure de concentration spatiale, le Gini ne tient aucun compte de la proximité dans l'espace des différentes zones de forte densité.

LA VALEUR MAXIMUM DU COEFFICIENT GINI



Zone	v(i)	w(i)	w(i)/v(i)	Cv(i)	Cw(i)	Cv(i)-Cw(i)	|v(i)-w(i)|
A	0,20	0,00	0,00	0,20	0,00	0,20	0,20
B	0,20	0,00	0,00	0,40	0,00	0,40	0,20
C	0,20	0,00	0,00	0,60	0,00	0,60	0,20
D	0,20	0,00	0,00	0,80	0,00	0,80	0,20
E	0,20	1,00	5,00	1,00	1,00	0,00	0,80
Total	1,00	1,00					1,60

Indice de dissimilarité (*D*) = 0,80

Coefficient Gini = 0,80

Source : Lorenz2.wb3, Lorenz3 et 4

EXEMPLE NUMÉRIQUE DE L'EFFET DE L'AGRÉGATION

Données initiales (« détaillées »)

	Superf.	Population		Densité	
		Période 0	Période *t*	Période 0	Période *t*
Zone 1	1	10	80	10	80
Zone 2	1	80	10	80	10
Zone 3	1	10	10	10	10

$G_0 = G_t = 0{,}47$, même si le centre de gravité de la population s'est déplacé vers la Zone 1.

Agrégation des zones 2 et 3 (découpage A)

	Superf.	Population		Densité	
		Période 0	Période *t*	Période 0	Période *t*
Zone 1	1	10	80	10	80
Zones 2 et 3	2	90	20	45	10

$G'_0 = 0{,}23$; $G'_t = 0{,}47$

$G'_t = G_t = 0{,}47$, puisque les zones agrégées sont d'égale densité (spécificité) à la période *t*.

Agrégation des zones 1 et 2 (découpage B)

	Superf.	Population		Densité	
		Période 0	Période *t*	Période 0	Période *t*
Zones 1 et 2	2	90	90	45	45
Zone 3	1	10	10	10	10

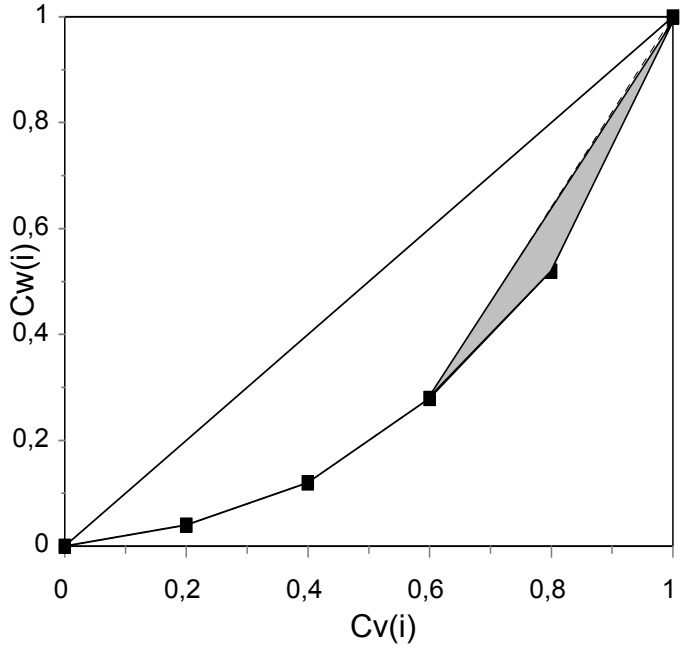
$G''_0 = G''_t = 0{,}23 < G_0 = G_t = 0{,}47$

Conclusions

- Sensibilité au découpage : les résultats « détaillés », ceux du découpage A et ceux du découpage B sont différents.
- Effet de l'agrégation : la valeur de l'indice de Gini diminue lorsqu'on agrège, **sauf** si on agrège des catégories (zones) de même spécificité (densité).

Source : Lorenz2.wb3, Lorenz3 et 4

EFFET DE L'AGRÉGATION SUR LE COEFFICIENT GINI



Zona	*v*(*i*)	*w*(*i*)	*w*(*i*)/*v*(*i*)	*Cv*(*i*)	*Cw*(*i*)	*Cv*(*i*)−*Cw*(*i*)	|*v*(*i*)−*w*(*i*)|
A	0,20	0,04	0,2	0,20	0,04	0,16	0,16
B	0,20	0,08	0,4	0,40	0,12	0,28	0,12
C	0,20	0,16	0,8	0,60	0,28	0,32	0,04
D	0,20	0,24	1,2	0,80	0,52	0,28	0,04
E	0,20	0,48	2,4	1,00	1,00	0,00	0,28
Total	1,00	1,00					0,64
Agregation des catégories *D* et *E*							
D+*E*	0,40	0,72	1,80	1,00	1,00	0,00	0,32
Total	1,00	1,00					0,64

Indice de dissimilarité (D) = 0,32

Coefficient Gini = 0,416 avant l'agrégation

Coefficient Gini = 0,368 après l'agrégation

Source : Lorenz2.wb3, Lorenz3 et 4

DISTANCE ET DISSIMILARITÉ

Propriétés d'une fonction de distance :

(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 :

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

Distance euclidienne

$$d_e(a,b) = \sqrt{X_{ab}^2 + Y_{ab}^2}$$

où

$X_{ab} = |x_a - x_b|$

$Y_{ab} = |y_a - y_b|$

Distance rectilinéaire (métrique de Manhattan) :

$d_r(a,b) = X_{ab} + Y_{ab}$

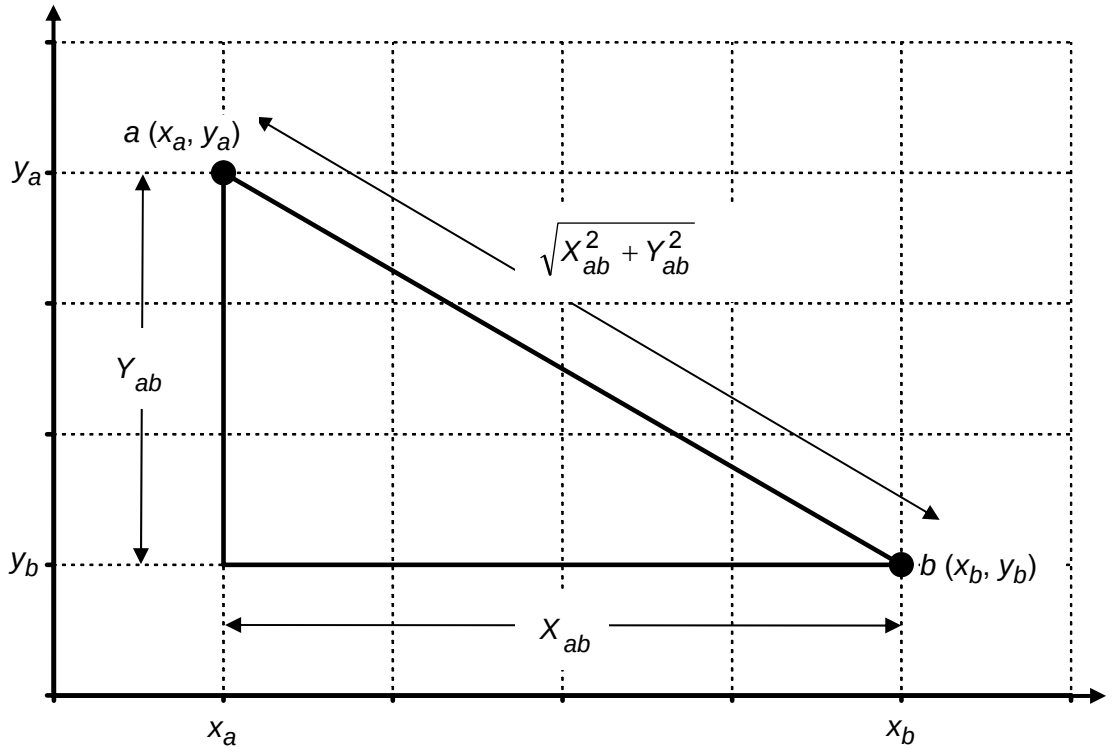
DISTANCES

Soit les points *a* et *b*, de coordonnées cartésiennes (x_a, y_a) et (x_b, y_b) repectivement.

Définissons

$X_{ab} = |x_a - x_b|$

$Y_{ab} = |y_a - y_b|$



Distance euclidienne

$$d_e(a,b) = \sqrt{X_{ab}^2 + Y_{ab}^2}$$

Distance rectilinéaire (métrique de Manhattan) :

$d_r(a,b) = X_{ab} + Y_{ab}$

DISTANCE ET MESURE DE LA DISSIMILARITÉ

La mesure de la distance est une mesure de la dissimilarité quant à la situation dans l'espace.

La situation dans un espace à 2 dimensions est décrite par 2 coordonnées :

	Latitude x	Longitude y
Point a	x_a	y_a
Point b	x_b	y_b
Différence	$x_a - x_b$	$y_a - y_b$

Définir une mesure de distance, c'est définir la façon de combiner les différences en une seule mesure.

La mesure de distance permet ensuite de déterminer, parmi les relations suivantes, lesquelles sont vraies :

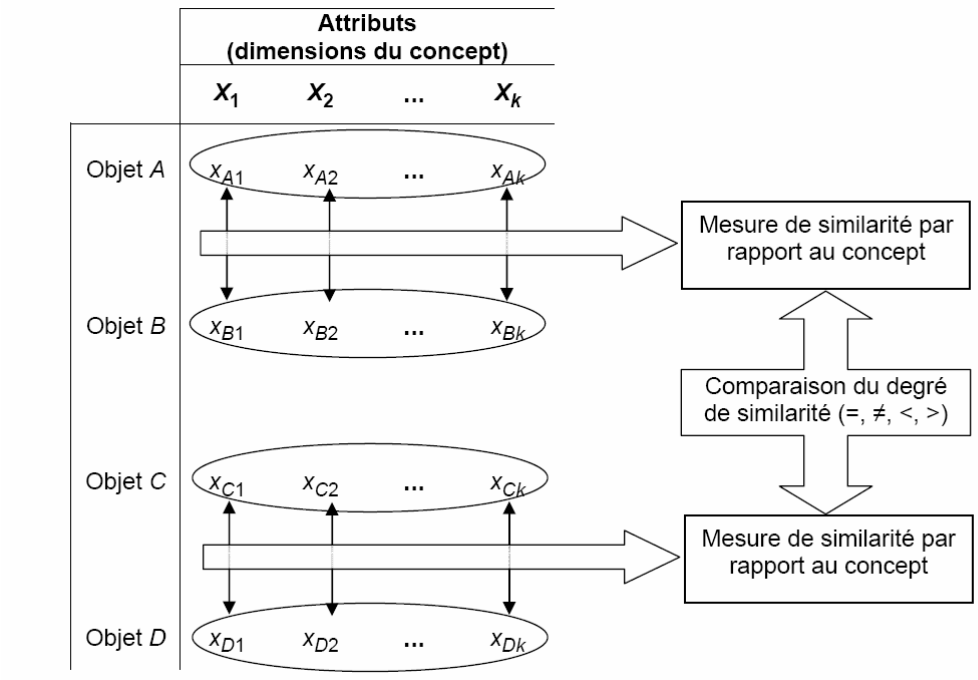
$d_r(a,b) = d_r(a,b)$

$d_r(a,b) \neq d_r(a,b)$

$d_r(a,b) > d_r(a,b)$

$d_r(a,b) < d_r(a,b)$

Mesure de la similarité



DISTANCES GÉNÉRALISÉES

Distances généralisées entre des distributions

Distributions comparées :

$p_{11}, p_{12}, ..., p_{1n}$

$p_{21}, p_{22}, ..., p_{2n}$

avec $\sum_i p_{ki} = 1$

- Distance rectilinéaire généralisée

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

  (formule similaire à celle de l'indice de dissimilarité *D*, mais sans la division par 2)

- Distance euclidienne généralisée

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

Distances généralisées entre des vecteurs d'attributs quelconques

Attributs des objets comparés :

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

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

Par exemple, une comparaison de quartiers d'une ville, caractérisés par...

x_{i1} = pourcentage de la population de moins de 15 ans
x_{i2} = taux de chômage
x_{i3} = revenu familial moyen
etc.

- Distance rectilinéaire généralisée

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

  (formule similaire à celle de l'indice de dissimilarité, mais sans la division par 2)

- Distance euclidienne généralisée

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

Mais si $\{x_{1i}\}$ et $\{x_{2i}\}$ ne sont pas des **distributions**, problème des **poids** (arbitraires ?)

Or le poids est fixé implicitement par les unités de mesure utilisées...

L'INDICE DE DISSIMILARITÉ (EXEMPLE NUMÉRIQUE)

Tableau de contingence : 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

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

Comparaison de la répartition géographique des branches *B*1 et *B*2

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

Mesure de la dissimilarité :

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

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

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

D = la moitié de la distance de Manhattan (distance rectilinéaire)

MESURE DE LA DISSIMILITUDE DANS UN TABLEAU 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 les branches dans toutes les zones
$p_{ij} = x_{ij} / x_{\bullet\bullet}$	fraction de l'emploi total global qui appartient à la branche *j* **et** qui se trouve 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 se trouve dans la zone *i*
$p_{j/i\bullet} = p_{ij} / p_{i\bullet}$	fraction de l'emploi total dans 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 se trouve dans la zone *i*

Dans l'exemple précédent, on mesure la dissimilitude entre

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

En général, 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 les distributions

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

PROPRIÉTÉS DE L'INDICE DE DISSIMILARITÉ

1. Remplit les conditions d'une mesure de distance (c'est la moitié de la distance rectilinéaire)
2. 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)
3. Domaine de variation (valeurs maximum et minimum)
   - $D = 0$ quand $v_i = w_i$ pour tout i (les deux distributions sont identiques)
   - $D = 1$ quand il y a ségrégation complète :
     soit $v_i > 0$, et alors, $w_i = 0$
     soit $w_i > 0$, et alors, $v_i = 0$
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. D est égal à l'écart vertical maximum entre la courbe de Lorenz et la diagonale.
6. 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.
7. 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.
8. Ne s'applique pas à des données négatives (ex. : comparaison des variations de l'emploi).

L'INDICE DE DISSIMILARITÉ ET LES PROPRIÉTÉS D'UNE MESURE DE DISTANCE

Propriétés d'une distance	**Indice de dissimilarité *D***
Non négativité : $d(a,b) \geq 0$	OUI
Identité : $d(a,b) = 0$ si, et seulement si $a = b$	OUI
Symétrie : $d(a,b) = d(b,a)$	OUI $D = \frac{1}{2}\sum_i \left\|v_i - w_i\right\| = \frac{1}{2}\sum_i \left\|w_i - v_i\right\|$
inégalité triangulaire : $d(a,c) \leq d(a,b) + d(b,c)$	OUI

Normal : *D* est la demie de la distance rectilinéaire généralisée (distance de Manhattan)

L'INDICE DE DISSIMILARITÉ ET LES PROPRIÉTÉS D'UNE MESURE D'INÉGALITÉ

Propriétés d'une mesure d'inégalité	**Indice de dissimilarité *D***
Une mesure d'inégalité doit prendre des valeurs non négatives.	OUI
Une mesure d'inégalité doit prendre la valeur zéro si, et seulement si, la distribution observée est identique à la distribution de référence.	OUI
Toutes les observations doivent être traitées de la même manière.	OUI
Mesure indépendante de la valeur moyenne de la variable examinée ou de la taille de la population dont on étudie la distribution.	OUI, puisque *D* est calculé à partir de la distribution.
L'agrégation d'observations affichant le même degré de spécificité ne doit pas changer la valeur de la mesure.	OUI
Principe de transfert de Pigou-Dalton	NON

L'INDICE DE DISSIMILARITÉ

EXEMPLE DE SÉGRÉGATION TOTALE

Indice de dissimilarité

ETHNIE	Nombres Martiens	Nombres Terriens	Nombres Total	Répartitions Martiens v_i	Répartitions Terriens w_i	Répartitions Total	Écart $\lvert v_i - w_i \rvert$
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 :

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

Ainsi, *D* varie entre 0 et 1

Et voilà pourquoi on divise par 2 !

INDICE DE DISSIMILARITÉ

VALEUR MAXIMUM

Démonstration que $D = 1$ lorsqu'il y a ségrégation complète

SOIT $v_i = 0$, et alors $\lvert v_i - w_i \rvert = \lvert 0 - w_i \rvert = w_i = 0 + w_i = v_i + w_i$

SOIT $w_i = 0$, et alors $\lvert v_i - w_i \rvert = \lvert v_i - 0 \rvert = v_i = v_i + 0 = v_i + w_i$

Il s'ensuit

$$D^{\max} = \frac{1}{2}\sum_i \lvert v_i - w_i \rvert = \frac{1}{2}\sum_i (v_i + w_i)$$

$$D^{\max} = \frac{1}{2}\left(\sum_i v_i + \sum_i w_i\right) = \frac{1+1}{2} = 1$$

INTERPRÉTATION MÉTAPHORIQUE
RENDRE LA DISTRIBUTION *B*2 IDENTIQUE À *B*1
(EXEMPLE NUMÉRIQUE)

Comparaison de la répartition géographique

des branches *B*1 et *B*2

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

« Excédents » de *B*2 sur *B*1 :

$= 0{,}142 + 0{,}083 = 0{,}225$ (*Z*1 et *Z*2)

« Déficits » de *B*2 par rapport à *B*1 :

$= 0{,}225$ (*Z*3)

Interprétation métaphorique :

« Il faut prendre 22,5 % (= *D*) des emplois de *B*2, dont 14,2 % dans *Z*1 et 8,3 % dans *Z*2 et il faut les déplacer vers *Z*3 ».

Ou, réciproquement :

« Il faut prendre 22,5 % des emplois de *B*1 dans *Z*3 ("excédentaires") et les déplacer vers les autres zones : 14,2 % dans *Z*1 et 8,3 % dans *Z*2 ».

Ou, en nombre d'emplois :

- Si on déplace les emplois de *B*2, ce sont 22,5 % de 600 emplois = 135 emplois.
- Si on déplace les emplois de *B*1, ce sont 22,5 % de 120 emplois = 27 emplois.

Mais il ne faut pas prendre la métaphore au pied de la lettre !

L'INDICE DE DISSIMILARITÉ ET LA COURBE DE LORENZ
ÉCART MAXIMUM ENTRE LA COURBE ET LA DIAGONALE

L'écart entre la courbe de Lorenz et la diagonale est donné par $Cv_k - Cw_k$

Pour quel *k* atteint-on la valeur maximum de $Cv_k - Cw_k$?

Pour chaque *k*, on a $Cv_k - Cw_k = \sum_{i=1}^{k} v_i - \sum_{i=1}^{k} w_i = \sum_{i=1}^{k} (v_i - w_i)$

Lorsque les observations sont en ordre croissant de spécificité, on a

$$\frac{w_1}{v_1} < \frac{w_2}{v_2} < \cdots < \frac{w_n}{v_n}$$

Donc, pour les premières observations, $v_i \geq w_i$ et pour les dernières, $w_i \geq v_i$

Par conséquent, tant que $v_i \geq w_i$, $Cv_i - Cw_i \geq Cv_{i-1} - Cw_{i-1}$

Pour trouver le maximum, il suffit de n'additionner que les valeurs positives (qui viennent toutes avant les négatives) : $MAX_k \left[Cv_k - Cw_k \right] = \sum_{\substack{i \text{ lorsque} \\ v_i > w_i}} (v_i - w_i)$

Mais puisque $\sum_{i=1}^{n} (v_i - w_i) = 0$, on a $\sum_{\substack{i \text{ lorsque} \\ v_i > w_i}} (v_i - w_i) = - \sum_{\substack{i \text{ lorsque} \\ v_i < w_i}} (v_i - w_i)$

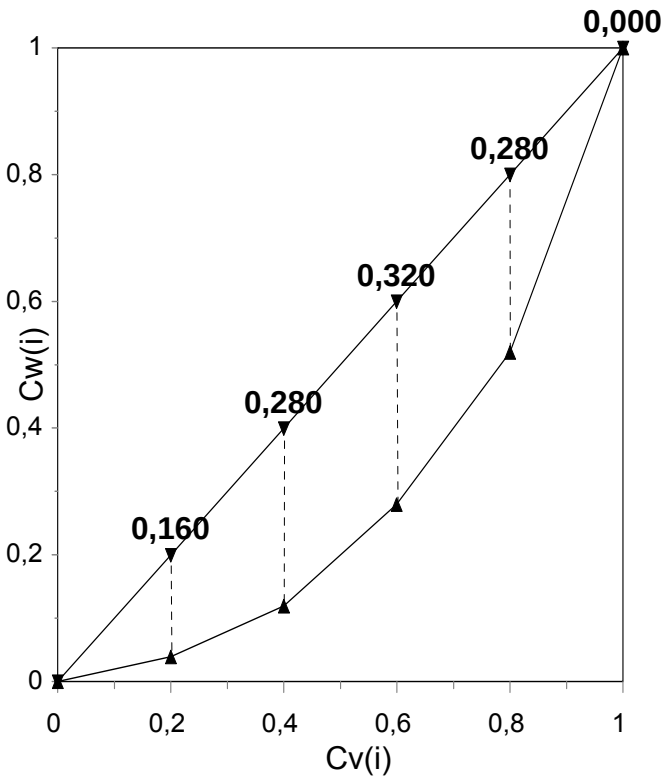
De plus, $\sum_{\substack{i \text{ lorsque} \\ v_i > w_i}} (v_i - w_i) - \sum_{\substack{i \text{ lorsque} \\ v_i < w_i}} (v_i - w_i) = \sum_i |v_i - w_i|$,

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

Donc,

$$MAX_k \left[Cv_k - Cw_k \right] = \sum_{\substack{i \text{ lorsque} \\ v_i > w_i}} (v_i - w_i) = \frac{1}{2} \sum_i |v_i - w_i| = D$$

ÉCART ENTRE LA COURBE DE LORENZ ET LA DIAGONALE



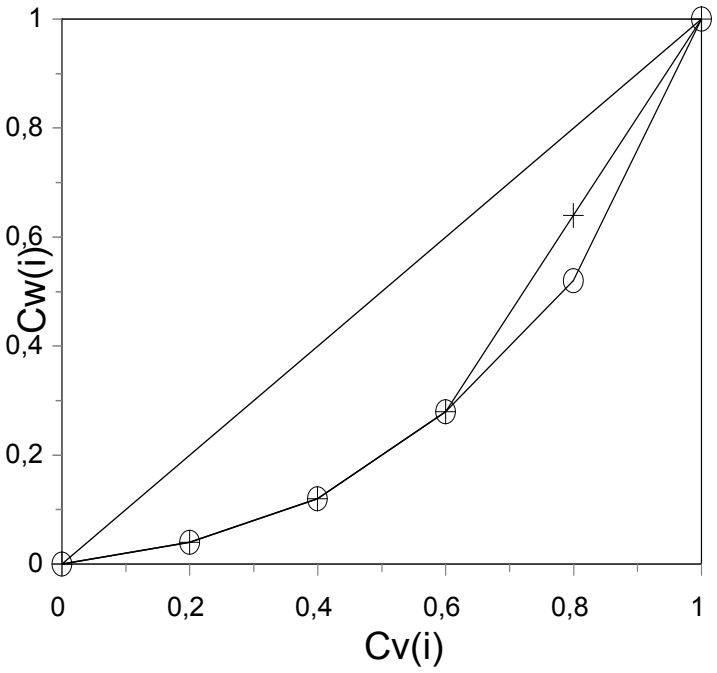
Zone	v(i)	w(i)	w(i)/v(i)	Cv(i)	Cw(i)	Cv(i)-Cw(i)	|v(i)-w(i)|
A	0,20	0,04	0,60	0,20	0,04	0,16	0,16
B	0,20	0,08	0,90	0,40	0,12	0,28	0,12
C	0,20	0,16	0,92	0,60	0,28	0,32	0,04
D	0,20	0,24	0,96	0,80	0,52	0,28	0,04
E	0,20	0,48	1,14	1,00	1,00	0,00	0,28
Total	1,00	1,00					0,64

Indice de dissimilarité D = 0,32

Coefficient Gini = 0,416

Source : Lorenz2.wb3, Lorenz1 et 5

INSENSIBILITÉ DE L'INDICE DE DISSIMILARITÉ À CERTAINS CHANGEMENTS



Distribution «O»

Zone	$v(i)$	$w(i)$	$w(i)/v(i)$	$Cv(i)$	$Cw(i)$	$Cv(i)–Cw(i)$	$\|v(i)–w(i)\|$
A	0,20	0,04	0,2	0,20	0,04	0,16	0,16
B	0,20	0,08	0,4	0,40	0,12	0,28	0,12
C	0,20	0,16	0,8	0,60	0,28	0,32	0,04
D	0,20	0,24	1,2	0,80	0,52	0,28	0,04
E	0,20	0,48	2,4	1,00	1,00	0,00	0,28
Total	1,00	1,00					0,64

Distribution «+»

Zone	$v(i)$	$w(i)$	$w(i)/v(i)$	$Cv(i)$	$Cw(i)$	$Cv(i)–Cw(i)$	$\|v(i)–w(i)\|$
A	0,20	0,04	0,60	0,20	0,04	0,16	0,16
B	0,20	0,08	0,90	0,40	0,12	0,28	0,12
C	0,20	0,16	0,92	0,60	0,28	0,32	0,04
D	0,20	0,36	1,80	0,80	0,64	0,16	0,16
E	0,20	0,36	1,80	1,00	1,00	0,00	0,16
Total	1,00	1,00					0,64

Indice de dissimilarité $D = 0{,}32$

Indice de Gini = 0,416 pour la distribution «O»

Indice de Gini = 0,368 pour la distribution «+»

Source : Lorenz2.wb3, Lorenz1 et 5

L'INDICE DE DISSIMILARITÉ COMME MESURE DE LA CONCENTRATION DE LA POPULATION

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 : 0,236

LE COEFFICIENT DE LOCALISATION N'EST PAS L'INDICE DE DISSIMILARITÉ

Bien qu'ils se calculent de la même manière,

l'indice de dissimilarité et le coefficient de localisation sont différents !

Emploi par zone et par branche

BRANCHE	$B1$	$B2$	$B3$	$B1 + B2$	Total
ZONE					
$Z1$	48	325	287	373	660
$Z2$	27	185	148	212	360
$Z3$	45	90	45	135	180
Total	120	600	480	720	1200

Distribution de l'emploi entre zones

BRANCHE	$B1$	$B2$	$B3$	$B1 + B2$	Total
ZONE					
$Z1$	0,400	0,542	0,598	0,518	0,550
$Z2$	0,225	0,308	0,308	0,294	0,300
$Z3$	0,375	0,150	0,094	0,188	0,150
Total	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	$B3$	Total	Dif.absol.	$B1 + B2$	Dif.absol.
ZONe					
$Z1$	0,598	0,550	0.048	0,518	0,080
$Z2$	0,308	0,300	0.008	0,294	0,014
$Z3$	0,094	0,150	0.056	0,188	0,094
Total	1,000	1,000	0,113	1,000	0,188

Mesure de la dissimilarité :

Indice de dissimilarité $D = \dfrac{|0{,}080| + |0{,}014| + |-0{,}094|}{2} = 0{,}094$

Coef. de localisation $CL = \dfrac{|0{,}048| + |0{,}008| + |-0{,}056|}{2} = 0{,}056$

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

APPLICATION DE L'INDICE DE DISSIMILARITÉ À UNE DICHOTOMIE

Bien qu'ils se calculent de la même manière,

l'indice de dissimilarité et le coefficient de localisation sont différents !

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

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

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

Démonstration :

D est appliqué à une dichotomie. Donc

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

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

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

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

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

L'INDICE DE DISSIMILARITÉ ET COEFFICIENT DE LOCALISATION

EXEMPLE DE SÉGRÉGATION TOTALE

Indice de dissimilarité

ETHNIE	Nombres			Répartitions			Écart
	Martiens	Terriens	Total	Martiens	Terriens	Total	
	x_{i1}	x_{i2}	$x_{i1} + x_{i2}$	$p_{i/\bullet 1}$	$p_{i/\bullet 2}$	$p_{i\bullet}$	$\lvert p_{i/\bullet h} - p_{i/\bullet k} \rvert$
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 :

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

Coefficient de localisation

ETHNIE	Nombres		Répartitions		Écart
	Martiens	Total	Martiens	Total	
	x_i	y_i	v_i	w_i	$\lvert v_i - w_i \rvert$
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 :

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

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

Indice de discrimination

Indice de discrimination :

$$\frac{\left(\frac{0{,}40 + 0{,}13 + 0{,}23 + 0{,}30}{2}\right)}{0{,}53} = 1{,}00$$