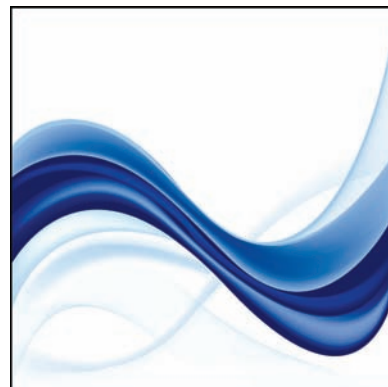


## Article

# La courbe concentration-couverture : un outil pour les études écologiques

par Philippe Finès

Décembre 2008



# La courbe concentration-couverture : un outil pour les études écologiques

par *Philippe Finès*

## Résumé

### Contexte

Une approche sélective peut être utilisée dans une étude écologique lorsqu'on cherche à sélectionner un sous-ensemble d'unités d'analyse (UA) et à tirer des conclusions au sujet d'une population d'intérêt (PI) uniquement d'après ces UA. Les résultats pour la PI seront fiables si cette population est concentrée dans les UA sélectionnées et rares dans les autres. Cet article décrit un outil graphique permettant de déterminer si ces conditions sont satisfaites.

### Données et méthodes

Des données sur les populations d'origine inuite et métisse tirées du Recensement de la population du Canada de 1996 sont utilisées à titre d'illustration. À partir d'un tableau des statistiques de classification, il est possible de créer une courbe concentration-couverture pour une PI donnée. La forme de la courbe indique s'il est possible de choisir un seuil tel que la concentration et la couverture de la PI seront toutes deux adéquates.

### Résultats

La courbe concentration-couverture montre que, parmi les peuples autochtones vivant dans les régions rurales, la population inuite est classable, mais non la population métisse.

### Interprétation

Cette méthode peut être appliquée à toute étude écologique dont le but est de déterminer la proportion d'individus ayant en commun une caractéristique unique définie par une variable binaire.

## Mots-clés

Classification, démographie, études écologiques, répartition de la population.

## Auteur

Philippe Finès (613-951-3896; [Philippe.Fines@statcan.gc.ca](mailto:Philippe.Fines@statcan.gc.ca)) travaille à la Division de l'information et de la recherche sur la santé, à Statistique Canada, Ottawa (Ontario) K1A 0T6.

Dans les études écologiques, les sujets sont regroupés en unités d'analyse (UA) au lieu d'être considérés séparément. Les résultats peuvent alors être exprimés en termes de proportions d'individus qui sont membres d'une population d'intérêt (PI) dans ces UA. Une approche sélective est utilisée quand on cherche à choisir un sous-ensemble d'UA et à tirer des conclusions au sujet d'une PI en se fondant uniquement sur ces UA. Si les UA sélectionnées ne contiennent que des individus appartenant à la PI, et si les individus formant la PI ne sont présents que dans ces UA, l'approche sélective est de toute évidence valide, dans la mesure où elle produit des résultats d'après lesquels il est possible de tirer des conclusions au sujet de la PI.

Toutefois, dans la plupart des cas, il n'est pas facile de déterminer si l'approche sélective est valide, et il est important de savoir si les UA sélectionnées représentent adéquatement la PI, c'est-à-dire si la PI est fortement concentrée dans ces UA et rare dans celles qui n'ont pas été sélectionnées. Nous proposons dans le présent article une méthode permettant de déterminer si ces conditions sont satisfaites.

Nous décrivons en détail la méthode proposée et illustrons comment elle s'applique aux études des peuples autochtones au Canada, en particulier les Inuits et les Métis vivant dans les régions rurales. Nous la comparons ensuite à des méthodes de classification apparentées et discutons d'autres considérations.

## Définitions

Dans les situations où les UA sélectionnées sont connues,

- la « proportion » (pour une UA donnée) est définie par le nombre d'individus de la PI divisé par le nombre total d'individus;
- la « concentration » est le nombre d'individus de la PI divisé par le nombre total d'individus dans les UA sélectionnées;
- la « couverture » est le nombre d'individus de la PI dans les UA sélectionnées divisé par le nombre total d'individus de la PI.

Partant de ces concepts, nous dirons qu'une PI est « classable » (et par conséquent, que l'approche sélective est valide) si nous pouvons choisir un seuil  $\alpha$  (compris entre 0 et 1) tel que, en sélectionnant uniquement les UA dans lesquelles la PI représente une proportion au moins égale à  $\alpha$ , nous

obtenons une concentration ainsi qu'une couverture élevées de la PI. À titre d'illustration, nous considérons comme PI deux groupes autochtones, à savoir les Inuits et les Métis, et comme UA, les secteurs de dénombrement (SD) du recensement.

### Étapes préliminaires

La méthode que nous appliquons pour déterminer si une PI est classable est basée sur la liste complète des UA, triées par ordre croissant de la proportion de leur population constituée d'individus de la PI. Chaque ligne de la liste correspond à une UA et indique le nombre d'individus de la PI dans l'UA, la population totale de l'UA et la proportion du total qui sont des individus de la PI. (Comme il existe 42 926 secteurs de dénombrement au Canada, il n'est pas possible d'inclure la liste complète dans le présent article,

mais elle peut être obtenue sur demande auprès de l'auteur.) La procédure est la suivante :

- Choisissons une ligne,  $i$ , de la liste.
- Additionnons les nombres d'individus de la PI à partir de la ligne  $i$  jusqu'au bas de la liste, puis divisons cette somme par le nombre total d'individus de la PI. Nous obtenons ainsi la *couverture* de la PI donnée par toutes les UA dans lesquelles la proportion de la PI est au moins égale à celle figurant à la ligne  $i$ .
- Divisons le même numérateur (le nombre d'individus de la PI de la ligne  $i$  jusqu'au bas de la liste) par le nombre total d'individus dans les UA de la ligne  $i$  jusqu'au bas de la liste. Nous obtenons ainsi la *concentration*, déterminée par toutes les UA dans lesquelles la proportion de la PI est au moins égale à celle de la ligne  $i$ .

**Tableau 1**  
**Statistiques de classification<sup>†</sup>, population d'origine inuite**

Seuil $\alpha$	Nombre de secteurs de dénombrement (SD) sélectionnés (SD où la proportion d'individus de la population d'intérêt est $\geq \alpha$ )	Nombre total d'individus dans les SD sélectionnés	Nombre d'individus dans les SD sélectionnés qui font partie de la population d'intérêt	Proportion d'individus dans les SD sélectionnés qui font partie de la population d'intérêt = Concentration	Nombre total d'individus dans les SD sélectionnés qui font partie de la population complémentaire	Proportion d'individus dans les SD sélectionnés qui font partie de la population complémentaire	Couverture = Sensibilité	Spécificité	Complément de la spécificité
(1)	(2)	(3)	(4) = A	(5) = (4)/(3)	(6) = (3)-(4) = B	(7) = (6)/(3)	(8) = A/(A+C)	(9) = B/(B+D)	(10) = D/(B+D)
0,00	42 926	28 454 565=N	39 845 = A+C	0,00140	28 414 720=B+D	0,99860	1,00000	0,00000	1,00000
0,05	91	54 265	35 085	0,64655	19 180	0,35345	0,88054	0,99932	0,00068
0,10	75	48 230	34 665	0,71874	13 565	0,28126	0,87000	0,99952	0,00048
0,15	68	45 135	34 300	0,75994	10 835	0,24006	0,86084	0,99962	0,00038
0,20	65	42 885	33 875	0,78990	9 010	0,21010	0,85017	0,99968	0,00032
0,25	64	41 605	33 570	0,80687	8 035	0,19313	0,84251	0,99972	0,00028
0,30	63	40 715	33 335	0,81874	7 380	0,18126	0,83662	0,99974	0,00026
0,35	62	40 145	33 140	0,82551	7 005	0,17449	0,83172	0,99975	0,00025
0,40	61	39 920	33 055	0,82803	6 865	0,17197	0,82959	0,99976	0,00024
0,45	58	37 840	32 170	0,85016	5 670	0,14984	0,80738	0,99980	0,00020
0,50	56	37 005	31 775	0,85867	5 230	0,14133	0,79747	0,99982	0,00018
0,55	56	37 005	31 775	0,85867	5 230	0,14133	0,79747	0,99982	0,00018
0,60	54	35 990	31 200	0,86691	4 790	0,13309	0,78303	0,99983	0,00017
0,65	51	33 750	29 790	0,88267	3 960	0,11733	0,74765	0,99986	0,00014
0,70	50	33 380	29 540	0,88496	3 840	0,11504	0,74137	0,99986	0,00014
0,75	46	27 975	25 585	0,91457	2 390	0,08543	0,64211	0,99992	0,00008
0,80	43	26 555	24 480	0,92186	2 075	0,07814	0,61438	0,99993	0,00007
0,85	43	26 555	24 480	0,92186	2 075	0,07814	0,61438	0,99993	0,00007
0,90	35	21 600	20 110	0,93102	1 490	0,06898	0,50471	0,99995	0,00005
0,95	4	1 275	1 220	0,95686	55	0,04314	0,03062	1,00000	0,00000
1,00	0	0	0	...	0	...	0,00000	1,00000	0,00000

<sup>†</sup> Les nombres compris entre 0 et 1 sont exprimés avec cinq décimales

... n'ayant pas lieu de figurer

Nota : Les lettres correspondent aux cellules de la matrice de classification (tableau 2).

Source : Recensement du Canada de 1996.

• Considérons une autre ligne, j, de la liste tel que la proportion de la PI y soit plus élevée qu'à la ligne i (autrement dit, la ligne j se trouve plus bas dans la liste). Par construction, la *couverture* est plus faible dans la ligne j que dans la ligne i, mais la *concentration* y est plus élevée.

**Tableau des statistiques de classification**

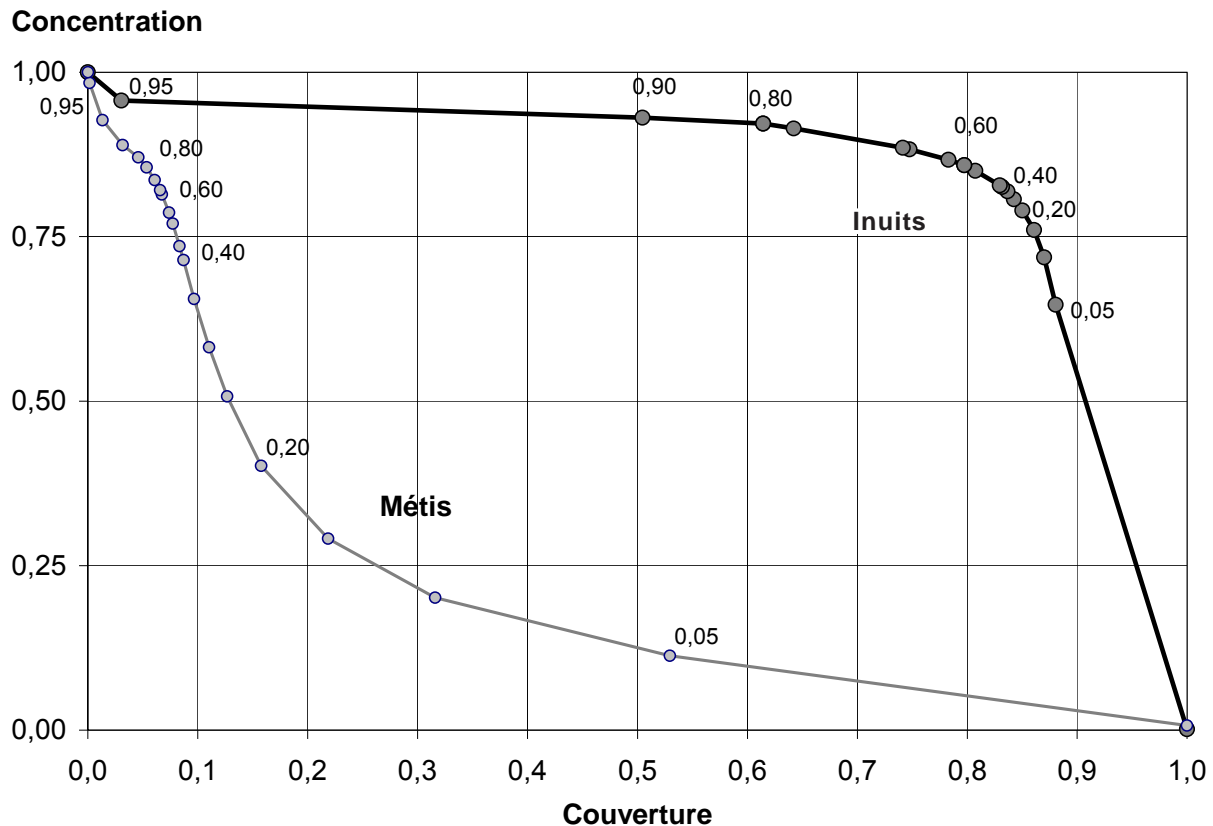
Partant de la liste des UA, nous pouvons produire le tableau des statistiques de classification. Ce tableau (présenté pour la population inuite au tableau 1) est tiré de la liste des UA (dont seules les proportions qui sont des multiples de 0,05 sont présentées). Les lignes correspondent aux seuils 0,00, 0,05, 0,10, ..., 1,00. (Nous incluons la ligne

pour le seuil 0,00 par souci de complétude, mais il ne s'agit pas d'une valeur de seuil utilisable.) Les colonnes (1) à (5) contiennent, respectivement, le seuil  $\alpha$ , le nombre d'UA sélectionnées (c.-à-d., le nombre d'UA dont la proportion de la PI est au moins égale au seuil), le nombre total d'individus dans les UA sélectionnées, le nombre d'individus de la PI dans les UA sélectionnées et la concentration de la PI dans les UA sélectionnées. Les colonnes (6) et (7) concernent la population complémentaire et la colonne (8) donne la couverture de la PI. Nous examinerons plus loin les colonnes (9) et (10), de même qu'une autre interprétation de la colonne (8); nous donnerons aussi une explication des lettres A, B, C, D et N.

À mesure que le seuil augmente, le nombre d'UA sélectionnées diminue. Par exemple, si nous fixons le seuil à 25 %, nous sélectionnerons l'ensemble des UA dans lesquelles la PI représente au moins 25 % de la population. L'utilisation d'un seuil plus élevé serait plus sélective : en effet, un sous-ensemble plus petit d'UA (qui sont toutes incluses dans le sous-ensemble précédent) serait alors sélectionné. Ce nouveau sous-ensemble contient un moins grand nombre d'individus appartenant à la PI, mais aussi un moins grand nombre d'individus qui n'en font pas partie.

Nous pouvons résumer ce raisonnement ainsi : si nous parcourons de haut en bas les lignes du tableau des statistiques de classification, le seuil augmente, ce qui réduit la couverture (en partant d'un maximum

**Figure 1**  
**Courbe concentration-couverture, populations inuite et métisse vivant dans les régions rurales, 1996**



Nota : Les points sur les courbes indiquent les valeurs de seuil ( $\alpha$ ).  
 Source : Recensement du Canada de 1996.

de 1) et accroît la concentration (jusqu'à une valeur inférieure ou égale à 1). Par conséquent, si le seuil est trop élevé, la couverture de la PI pourrait être insuffisante; s'il est trop faible, un grand nombre d'individus qui ne font pas partie de la PI pourraient être inclus, ce qui réduirait alors la concentration.

### Construction et interprétation de la courbe concentration-couverture

Pour qu'une PI soit classable, il doit être possible de choisir un seuil tel que la concentration et la couverture de cette PI soient toutes deux adéquates. Pour déterminer si ces critères sont satisfaits pour une PI donnée, nous représentons graphiquement la concentration en fonction de la couverture, ce qui nous donne la courbe concentration-couverture (courbe CC).

Les courbes CC pour les populations inuite et métisse sont présentées à la figure 1. Sur chaque courbe, chaque point représente une ligne particulière du tableau des statistiques de classification. Par souci de clarté, la valeur du seuil est indiquée pour certains points des courbes. (Le tableau des statistiques de classification pour la population des Métis n'est pas présenté dans l'article, mais peut être obtenu sur demande auprès de l'auteur.) Sur la courbe CC de la population inuite, pour deux paires de seuils ( $\alpha = 0,50, 0,55$  et  $\alpha = 0,80, 0,85$ ), le nombre de SD est le même pour les deux seuils, si bien que, dans chaque cas, les deux points de la paire ont la même couverture (0,80 et 0,61).

La partie supérieure gauche de la courbe CC de la population inuite s'étend presque horizontalement sur plus de la moitié de la longueur de l'axe des abscisses. Cette partie de la courbe représente les SD pour lesquels la concentration et la couverture sont élevées. Par conséquent, la forme de la courbe indique que la sélection de ces SD est suffisante pour obtenir une bonne classification de la population inuite. Par exemple, si l'on veut obtenir une couverture de 50 %, il suffit de

**Tableau 2**  
**Matrice de classification**

	Individus appartenant à la population d'intérêt	Individus n'appartenant pas à la population d'intérêt	Total
Résidents des unités d'analyse sélectionnées	A (« vrais positifs »)	B (« faux positifs »)	A+B
Résidents des unités d'analyse non sélectionnées	C (« faux négatifs »)	D (« vrais négatifs »)	C+D
Total	A+C	B+D	N

ne retenir que les SD dont la concentration est supérieure à environ 93 %; au tableau 1 et à la figure 1, ces SD sont ceux dans lesquels la proportion de la PI (c.-à-d. la population inuite) est supérieure ou égale à un seuil de 0,90. L'utilisation d'un seuil plus faible n'est pas nécessaire. (Pour une couverture de 75 %, on retiendra les SD donnant une concentration de 87 %, ce qui correspond à un seuil de 0,60.) Donc, selon les critères indiqués, la population inuite est une PI classable. Par exemple, le seuil utilisé par Wilkins et coll.<sup>1</sup> dans leur analyse de l'espérance de vie dans les régions habitées par les Inuits était de 33 %, ce qui a produit une concentration ainsi qu'une couverture d'environ 80 %.

Dans le cas des Métis, la situation est fort différente. Contrairement à la courbe obtenue pour la population inuite, la courbe CC de la population métisse a une forte pente à la gauche du graphique, puis s'aplatit par la suite (figure 1). La forme de la courbe indique qu'il est impossible de trouver un seuil produisant à la fois une forte couverture et une forte concentration de la population métisse : pour obtenir une couverture de 50 %, il faut que la concentration soit assez faible, soit 11 % (ce qui correspond à un seuil de 0,05); par ailleurs, pour une concentration d'environ 40 %, la couverture serait à peine de 16 % (avec un seuil de 0,20). Par conséquent, contrairement à la population inuite, celle des Métis ne

constitue pas une PI classable, même si elle est nettement plus nombreuse (199 235 contre 39 845).

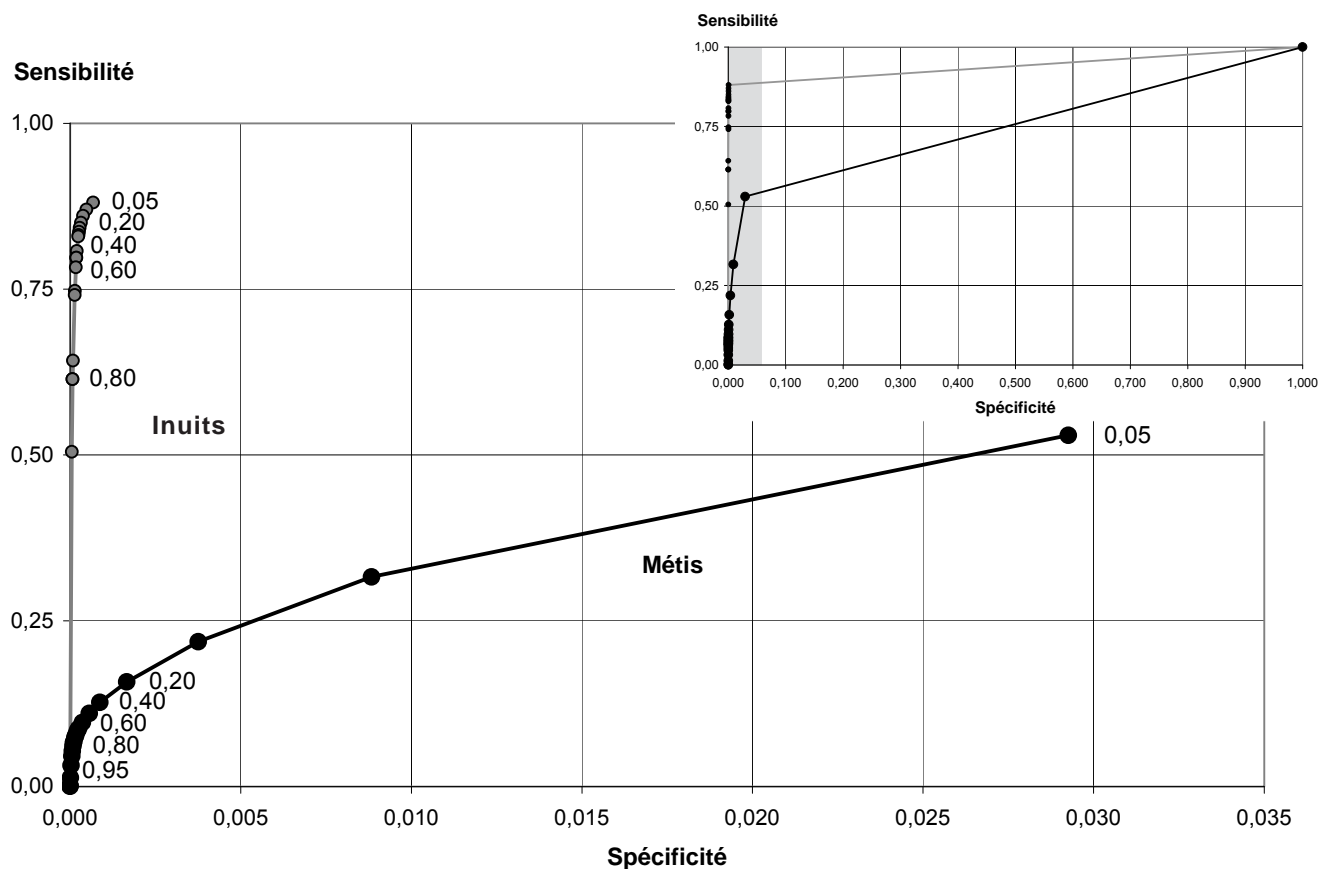
### Matrice de classification

Le tableau 1 montre que, quand le seuil augmente, le nombre d'UA sélectionnées diminue et le contenu des colonnes 3 (population totale dans les UA sélectionnées), 4 (nombre d'individus de la PI dans les UA sélectionnées) et 6 (nombre d'individus dans les UA sélectionnées qui n'appartiennent pas à la PI) diminue. Parallèlement, la proportion moyenne d'individus de la PI dans les UA augmente.

Nous pouvons expliquer ces résultats à l'aide des concepts de la théorie de la classification. Notre classification comporte deux variables binaires, les UA sélectionnées ou non sélectionnées, d'une part, et les individus qui sont membres de la PI ou qui ne le sont pas (population complémentaire), d'autre part. Nous pouvons utiliser une matrice de classification pour illustrer ces variables (tableau 2). Les N individus de la population canadienne sont répartis entre les quatre cellules de la matrice de la façon suivante :

- la cellule A contient les individus qui sont membres de la PI et qui vivent dans une UA sélectionnée, soit les « vrais positifs »;
- la cellule B contient les individus qui ne sont pas membres de la PI,

**Figure 2**  
**Courbe ROC, populations inuite et métisse vivant dans les régions rurales, 1996**



**Nota :** Le graphique complet correspond à la partie ombrée du graphique en cartouche. Les points sur les courbes indiquent les valeurs de seuil ( $\alpha$ ).

**Source :** Recensement du Canada de 1996.

- la cellule A contient les individus qui sont membres de la UA sélectionnée, soit les « vrais positifs »;
- la cellule C contient les individus qui sont membres de la PI, mais qui ne vivent pas dans une UA sélectionnée, soit les « faux négatifs »;
- la cellule D contient les individus qui ne sont pas membres de la PI et qui ne vivent pas dans une UA sélectionnée, soit les « vrais négatifs ».

Quand le seuil augmente, les valeurs de A et de B diminuent, tandis que celles de C et de D augmentent. Naturellement, les valeurs de A+C, B+D et N restent les mêmes. Donc, la sensibilité, qui est égale à  $A/(A+C)$  et est simplement un autre terme pour

désigner la couverture, diminue (colonne 8) tandis que la spécificité, qui correspond à  $D/(B+D)$ , augmente (colonne 9) et que le complément de la spécificité diminue (colonne 10).

### Comparaison à d'autres courbes

L'information de la courbe concentration-couverture peut être représentée graphiquement au moyen d'un autre type de courbe, à savoir la courbe ROC (de l'anglais *receiver operating characteristic*). Très souvent utilisée en classification<sup>2</sup>, la courbe ROC reflète la qualité d'une classification par rapport à tous les seuils choisis pour l'établir. Pour la construire, on

trace la sensibilité en fonction du complément de la spécificité. Quand le seuil augmente, la sensibilité et le complément de la spécificité diminuent (déplacement vers le bas et vers la gauche le long de la courbe ROC). Autrement dit, le coin inférieur gauche (coordonnées 0,0) correspond au seuil le plus élevé (1,00) et le coin supérieur droit (coordonnées 1,1), au seuil le plus faible.

La figure 2 donne les courbes ROC pour les populations inuite et métisse. Comme dans le cas des courbes CC, les valeurs de certains seuils sont indiquées à côté des points sur la courbe. Étant donné que les peuples autochtones représentent une faible proportion de la population du Canada (0,1 % pour

la population inuite, comme l'indique la ligne 0,00 du tableau 1), nous obtenons invariablement une très forte spécificité et, par conséquent, une très faible valeur pour le complément de la spécificité sur la courbe ROC. (C'est pourquoi la figure 2 comprend deux versions de la courbe ROC : la complète – en cartouche – et celle qui exclut le point de seuil 0.)

La courbe ROC constitue un critère de classification intuitif : la PI sera d'autant plus classable que la courbe s'approchera du coin supérieur gauche. À la figure 2, la courbe ROC de la population métisse est plus éloignée du coin supérieur gauche que celle de la population inuite, ce qui indique que la population métisse est moins bien classable que la population inuite.

Toutefois, la courbe CC fournit plus de renseignements que la courbe ROC, car elle offre un moyen visuel qualitatif de déterminer si une PI est classable. Une PI dont la courbe CC demeure à peu près horizontale sur une certaine distance dans le coin supérieur gauche du graphique est classable, tandis qu'une PI dont la courbe CC a une forte pente descendante à la gauche du graphique n'est pas classable. En outre, dans le cas de PI particulièrement rares, comme

les populations inuite et métisse, il est difficile d'afficher des valeurs de seuil particulières sur la courbe ROC, car les seuils les plus élevés sont entassés dans la partie inférieure gauche de la courbe. En revanche, sur la courbe CC, ces seuils sont dispersés.

Les courbes généralement utilisées pour mesurer l'inégalité, telles que la courbe de Lorenz<sup>3</sup> et la courbe d'inégalité du revenu de Kakwani<sup>4</sup>, contiennent une dimension supplémentaire (habituellement le revenu) et ne permettent pas de réaliser l'objectif proposé ici, qui est de déterminer si une PI est classable. La courbe CC ne s'appuie que sur une seule variable, la proportion de la PI. Son caractère unique tient au fait que la proportion est exprimée à l'aide de deux dénominateurs différents.

### Conclusion

Nous fournissons dans le présent article une définition opérationnelle d'une PI classable : il doit exister au moins un seuil pour lequel la concentration (proportion de la population des UA qui fait partie de la PI) et la couverture (proportion totale de la PI contenue dans les UA sélectionnées) sont toutes deux suffisamment élevées. Nous

donnons aussi une représentation visuelle de ce concept : si la PI est classable, la partie supérieure gauche de la courbe CC est à peu près horizontale.

Dans l'exemple choisi pour illustrer la méthode, la PI est définie selon une caractéristique ethnique (origine autochtone). Cependant, il est clair que la méthode peut être appliquée à toute étude écologique dont le but est de déterminer la proportion d'individus ayant en commun une caractéristique unique, surtout si cette proportion est faible. Cette caractéristique pourrait être, par exemple, une langue, un comportement ou une maladie. ■

### Remerciements

L'auteur remercie Russell Wilkins de ses commentaires au sujet d'une version antérieure de l'article, ainsi que Nancy Ross, Nancy Darcovich et deux réviseurs anonymes pour leurs commentaires.

## Références

1. R. Wilkins, S. Uppal, P. Finès *et al.*, « Espérance de vie dans les régions où vivent les Inuits au Canada, 1989 à 2003 », *Rapports sur la santé*, 19(1), 2008, p. 7-20 (Statistique Canada, n° 82-003 au catalogue).
2. J.F. Jekel, D.L. Katz et J.G. Elmore, *Epidemiology, Biostatistics and Preventive Medicine – Second edition*, Philadelphie, W. B. Saunders Company, 2001.
3. I. Kawachi et B.P. Kennedy, « The relationship of income inequality to mortality: Does the choice of indicator matter? », *Social Science and Medicine*, 45(7), 1997, p. 1121-1127.
4. N. Kakwani, A. Wagstaff et E. van Doorslaer, « Socioeconomic inequalities in health: Measurement, computation, and statistical inference », *Journal of Econometrics*, 77(1), 1997, p. 87-103.