Mesure de la biodiversité et de la structuration spatiale de l'activité économique par l'entropie

Eric Marcon, Florence Puech

15 décembre 2017

1 Introduction

Le terme « entropie » a été introduit par Clausius en 1865 pour sa nouvelle formulation du second principe de la thermodynamique énoncé par Carnot 40 ans plus tôt. Son étymologie grecque signifie transformation parce que le second principe concerne la variation d'entropie. Boltzmann a relié l'entropie d'un système au nombre possible de ses états en 1877. Shannon (1948) a enfin montré que le nombre d'états possibles d'un système était analogue au nombre de messages d'une longueur choisie pouvant être créés en assemblant les lettres d'un alphabet dont les fréquences des lettres sont fixées. La célèbre entropie de Shannon est, à une constante près, égale à celle de Boltzmann normalisée par la longueur du message, dont elle est indépendante. Cette propriété fondamentale lui permet de décrire la complexité d'un système non seulement par le nombre possible de ses états, mais plus simplement par la fréquence relative de ses composants. La théorie de l'information était née, les origines thermodynamiques rapidement oubliées, mais pas la physique statistique qui a généralisé les travaux de Boltzmann à des systèmes imparfaits (Tsallis, 1988).

L'entropie de Shannon est la base de toutes les mesures de diversité présentées ici. Son application à la diversité biologique, devenue biodiversité (Wilson & Peter, 1988), est fondamentale. Parallèlement, en économie, Theil (1967) développait des mesures d'inégalité et de concentration spatiales : l'indice de Theil est l'écart entre l'entropie de Shannon et son maximum possible, mais les développements méthodologiques ultérieurs sont restés en retrait de ceux de la mesure de la biodiversité.

L'objectif de cet article est transférer à la discipline de l'économie géographique les derniers développements de la mesure de la biodiversité pour compléter ses définitions de concentration spatiale et spécialisation, apporter des méthodes de mesure plus efficaces sur le plan empirique comme des estimateurs à biais réduit quand seulement un échantillon des données est disponible, et des solutions à une partie des problèmes classiques soulevés par ces approches, notamment celui la MAUP (Openshaw & Taylor, 1979, *Modifiable Areal Unit Problem*), c'est-à-dire la sensibilité des mesures à l'échelle d'observation. Les emprunts très nombreux de méthodes entre disciplines éloignées seront montrés.

2 La diversité définie comme quantité d'information

2.1 Entropie et théorie de l'information

Les textes fondateurs sont Davis (1941) et surtout Theil (1967) en économétrie, et Shannon (Shannon, 1948; Shannon & Weaver, 1963) pour la mesure de la diversité. Une revue est fournie par Maasoumi (1993).

Considérons une expérience dont les résultats possibles sont $\{r_1, r_2, \ldots, r_S\}$. La probabilité d'obtenir r_s est p_s, et $\mathcal{P} = \{p_1, p_2, \ldots, p_S\}$. Les probabilités sont connues *a priori*. Tout ce qui suit est vrai aussi pour des valeurs de r continues, dont on connaîtrait la densité de probabilité.

On considère maintenant un échantillon de valeurs de r. La présence de r_s dans l'échantillon est peu étonnante si p_s est grande : elle apporte peu d'information supplémentaire par rapport à la simple connaissance des probabilités. En revanche, si p_s est petite, la présence de r_s est surprenante. On définit donc une fonction d'information, $I(p_s)$, décroissante quand la probabilité augmente, de $I(0) = +\infty$ (ou éventuellement une valeur strictement positive finie) à $I(1) = 0$.

Chaque valeur observée dans l'échantillon apporte une certaine quantité d'information, dont la somme est l'information de l'échantillon. Patil & Taillie (1982) appellent l'information « rareté ».

La quantité d'information attendue de l'expérience est $\sum_{s=1}^{S} p_s I(p_s) = H(\mathcal{P})$. Si on choisit $I(p_s) = -\ln(p_s)$, $H(\mathcal{P})$ est l'indice de Shannon, mais bien d'autres formes de $I(p_s)$ sont possibles. $H(\mathcal{P})$ est appelée *entropie*. C'est une mesure de l'incertitude (de la volatilité) du résultat de l'expérience. Si le résultat est certain (une seule valeur p_S vaut 1), l'entropie est nulle. L'entropie est maximale quand les résultats sont équiprobables.

2.2 Application à la biodiversité

MacArthur (1955) est le premier à avoir introduit la théorie de l'information en écologie (Ulanowicz, 2001). MacArthur s'intéressait aux réseaux trophiques et cherchait à mesurer leur stabilité : l'indice de Shannon qui comptabilise le nombre de relations possibles lui paraissait une bonne façon de l'évaluer. Mais l'efficacité implique la spécialisation, ignorée par l'entropie qui est une mesure neutre (toutes les espèces y jouent le même rôle). MacArthur a abandonné cette voie.

Les premiers travaux consistant à généraliser l'indice de Shannon sont dus à Rényi (1961). La biodiversité sera abordée ici en termes d'espèces pour la simplicité de l'exposé, sans perte de généralité : les mêmes méthodes s'appliquent à la diversité génétique par exemple. L'entropie d'ordre q de Rényi utilise une fonction d'information paramétrique dans laquelle le paramètre q, librement choisi, donne une importance d'autant plus grande aux espèces rares qu'il est petit. L'entropie de Rényi d'ordre 0 est la richesse, c'est-à-dire le nombre d'espèces (précisément, ce nombre moins 1), l'entropie d'ordre 1 est celle de Shannon et l'entropie d'ordre 2, l'indice de diversité de Simpson (1949).

Hill (1973) transforme l'entropie de Rényi en *nombres de Hill*, qui en sont simplement l'exponentielle. Le souci de Hill était de rendre les indices de diversité intelligibles après l'article remarqué de Hurlbert (1971) intitulé « le non-concept de diversité spécifique ». Hurlbert reprochait à la littérature sur la diversité sa trop grande abstraction et son éloignement des réalités biologiques, notamment en fournissant des exemples dans lesquels l'ordre des communautés n'est pas le même selon l'indice de diversité choisi. Les nombres de Hill sont le nombre d'espèces équiprobables donnant la même valeur de diversité que la distribution observée.

Ces résultats avaient déjà été obtenus avec une autre approche par MacArthur (1965) et repris par Adelman (1969) dans la littérature économique.

Les nombres de Hill sont des « nombres effectifs » ou « nombres équivalents ». Le concept a été défini rigoureusement par Gregorius (1991), d'après Wright (1931) (qui avait le premier défini la taille effective d'une population) : étant donné une variable caractéristique (ici, l'entropie) fonction seulement d'une variable numérique (ici, le nombre d'espèces) dans un cas idéal (ici, l'équiprobabilité des espèces), le nombre effectif est la valeur de la variable numérique pour laquelle la variable caractéristique est celle du jeu de données.

2.3 Biais d'estimation

L'entropie est définie comme la somme pondérée sur toutes les espèces de l'information. Dans des systèmes très divers comme la forêt tropicale, inventorier la totalité des espèces est en général impossible. Estimer le nombre d'espèces total par le nombre d'espèces échantillonnées est évidemment incorrect : l'estimation du nombre d'espèces non observées a généré une abondante littérature (Chao, 2004).

Le problème concerne toutes les mesures de diversité : il n'est pas lié à un échantillonnage défaillant mais simplement à une variabilité inévitable et à l'impossibilité d'augmenter indéfiniment l'effort d'échantillonnage. Sa conséquence est une sous-estimation de la diversité, appelée « biais d'estimation » par Dauby & Hardy (2012).

Les espèces rares ont un rôle central dans le biais d'estimation parce qu'elles sont plus difficiles à observer. Les mesures de diversité qui leur donnent une grande importance (l'exemple le plus simple est la richesse spécifique) sont plus biaisées que les mesures qui ne prennent en compte que les espèces dominantes (comme l'indice de Simpson).

2.4 Entropie HCDT

Le physicien Tsallis (1988) propose une classe de mesures appelée entropie généralisée, définie par Havrda & Charvát (1967) pour la première fois (en cybernétique) et redécouverte plusieurs

fois, notamment par Daróczy (1970) (en théorie de l'information), d'où son nom *entropie HCDT* (voir Mendes *et al.* (2008), page 451, pour un historique complet).

Tsallis a montré que les indices de Simpson et de Shannon étaient des cas particuliers d'entropie généralisée. Ces résultats ont été complétés par d'autres et repris en écologie par Keylock (2005) et Jost (2006, 2007).

L'entropie HCDT est particulièrement attractive parce que sa relation avec la diversité au sens strict est simple, après introduction du formalisme adapté : celui des logarithmes déformés (Tsallis, 1994). Le logarithme déformé d'ordre q est une fonction définie de façon que l'entropie HCDT puisse être écrite comme l'entropie de Shannon, en le substituant au logarithme naturel : l'entropie d'ordre q est simplement le logarithme déformé de son nombre de Hill (Marcon *et al.*, 2014).

2.5 Phylodiversité

Les mesures neutres de la diversité considèrent que toutes les classes auxquelles les objets appartiennent sont différentes, sans que certaines soient plus différentes que d'autres. Par exemple, toutes les espèces sont équidistantes les unes des autres, qu'elles appartiennent au même genre ou à des familles différentes. Intuitivement, l'idée qu'une communauté de S espèces toutes de genres différents est plus diverse qu'une communauté de S espèces du même genre est satisfaisante.

Il s'agit donc de caractériser la différence entre deux classes d'objets, puis de construire des mesures de diversité en rapport (Pielou, 1975; May, 1990; Cousins, 1991). En écologie, ces différences sont fonctionnelles ou phylogénétiques, définissant la diversité fonctionnelle (Tilman *et al.*, 1997) ou la diversité phylogénétique (*phylodiversity*) (Webb *et al.*, 2006). Les premières propositions de ce type d'indices sont dues à Rao (1982).

Dans le cadre de la diversité phylogénétique traitée ici, les espèces sont placées dans un arbre représentant leur évolution. À partir de la racine de l'arbre (représentant l'ancêtre commun), une nouvelle période est définie à chaque ramification d'une branche quelconque jusqu'aux espèces présentes placées aux extrêmités des dernières branches. La longueur des branches représente le temps de l'évolution.

L'entropie HCDT est calculée à chaque période. L'entropie phylogénétique (Marcon *et al.*, 2014) est l'entropie moyenne calculée tout au long des périodes de l'arbre. Elle s'interprète comme l'information moyenne apportée par un individu observé au hasard, à une temps choisi au hasard. Elle peut être transformée en diversité de la même façon.

3 Diversité β et décomposition

La notion de diversité β a été introduite par (Whittaker, 1960, page 320) comme le niveau de changement dans la composition des communautés, ou le degré de différenciation des communautés, en relation avec les changements de milieu. La traduction de cette notion intuitive en une définition sans ambiguïté est encore une question de recherche et de débats. Anderson *et al.* (2011) fournissent une revue des analyses utiles de la diversité β en forme de guide à destination des écologues.

Trois niveaux de diversité sont définis : la diversité locale, ou α est la diversité moyenne de plusieurs communautés, définie comme précédemment. La diversité γ est celle de leur assemblage, mesurée de la même façon mais sans distinction de l'appartenance à telle ou telle communauté. Enfin, la diversité β, conceptuellement assez différente, mesure à quel point chacune des communauté est différente de l'assemblage Marcon *et al.* (2012).

La décomposition de la diversité relie ces trois mesures : l'entropie β est la différence entre les entropies γ et α : c'est l'information supplémentaire apportée par la connaissance de la composition de chaque communauté en plus de la connaissance de leur seul assemblage (Marcon *et al.*, 2014). La diversité β au sens strict est le rapport entre les diversités γ et α : il s'agit du nombre effectif de communautés, c'est à dire le nombres de communautés sans espèces communes qui fourniraient la même diversité β que les communautés observées.

Il est possible de décomposer hiérarchiquement la diversité (Pavoine *et al.*, 2016) au-delà de deux niveaux en considérant chaque communauté comme un assemblage de sous-communautés plus petites : la diversité α devient la diversité γ à une échelle plus détaillée.

3.1 Entropie et diversité

L'entropie est utile pour les calculs : la correction des biais d'estimation notamment. Elle mesure l'information moyenne apportée par la connaissance des fréquences relatives des espèces, ce qui est conceptuellement clair mais numériquement peu intuitif. Les nombres de Hill, ou *nombres équivalent d'espèces* ou *nombres d'espèces effectives* permettent une appréhension plus intuitive de la notion de biodiversité (Jost, 2006). En raison de leurs propriétés, notamment de décomposition, Jost (2007) les appelle « vraie diversité ». La diversité γ, nombre effectif d'espèces, se décompose en un nombre de communautés (diversité β) sans espèces communes possédant chacune le même nombre d'espèces équiprobables (diversité α).

L'intérêt de ces approches est de fournir une définition paramétrique de la diversité, qui donne plus ou moins d'importance aux espèces rares.

4 Transfert à l'économie géographique des méthodes de la biodiversité

Les recherches sur la structure spatiale de l'activité économique se sont principalement intéressées à la concentration spatiale, source d'externalités positives (Marshall, 1890; Weber, 1909; Krugman, 1991). La concentration spatiale va de pair avec la spécialisation (Houdebine, 1999; Cutrini, 2010). Le cadre conceptuel est le suivant : des employés peuvent être localisés dans une région quelconque d'un pays donné, et travailler dans un secteur économique quelconque. Les données sont le nombre d'employés de chaque secteur dans chaque région. Sous l'hypothèse nulle d'une distribution non structurée, la connaissance de la taille relative de chaque secteur et de chaque région donne l'espérance de ce nombre. La concentration spatiale d'un secteur économique mesurée par l'indice d'Ellison et Glaeser (Ellison & Glaeser, 1997) est l'écart entre la part de chaque région dans ce secteur et la taille relative des régions. De façon symétrique, la spécialisation d'une région peut être définie comme l'écart de la distribution des poids relatifs de ses secteurs économiques à leurs poids dans l'ensemble du pays. Les deux peuvent être combinés pour définir une mesure de diversité jointe (Gregorius, 2010), écart entre la distribution des couples secteur $\times$ région et leur valeur attendue en absence de structuration. Cutrini (2010) a défini cette diversité jointe, mesurée par l'entropie de Shannon, comme un « indice de localisation globale ».

Les développements méthodologiques du domaine de la diversité peuvent être appliqués à ce cadre pour généraliser cette mesure de localisation globale à l'entropie HCDT. La spécialisation est l'écart entre la diversité et sa valeur maximale. La diversité économique d'une région peut être calculée par les méthodes de la biodiversité, donnnant un poids arbitraire aux secteurs de petite taille. La concentration spatiale est la même mesure du point de vue des secteurs plutôt que des régions, permettant l'application des mêmes méthodes en transposant simplement la matrice des données.

Les problèmes classiques de sensibilité des mesures de concentration spatiale et de spécialisation en espace discret (Arbia, 1989; Openshaw & Taylor, 1979) peuvent être largement réduits en considérant l'emboîtement des échelles spatiales ou des secteurs économiques plus ou moins agrégés de la même façon qu'une phylogénie. À titre d'exemple, le problème d'échelle est l'incohérence des mesures de concentration géographique considérées à des échelles différentes. Sa résolution théorique est immédiate dans le cadre de la décomposition de la diversité présentée dans la section précédente : la concentration (ou la spécialisation) à un niveau agrégé (par exemple un pays) est égale à sa valeur moyenne au niveau désagrégé (par exemple les régions de ce pays) à laquelle s'ajoute la divergence entre régions (analogue à la diversité β) dont l'ignorance est le fondement du problème.

D'autre part, les mesures de concentraton spatiale dépendent du niveau d'agrégation des secteurs économiques auxquelles elles sont appliquées, au point de fausser les comparaisons entre secteurs (Bickenbach *et al.*, 2013). L'indépendance est une des propriétés des mesures de concentration spatiales requises par Combes & Overman (2004). La diversité phylogénétique est un outil pour régler ce problème en prenant en compte l'arborescence complète de la classification sectorielle.

L'article détaillera ces mesures et les illustrera par un exemple traité en détail.

Références

Adelman, M.A. (1969) Comment on the "H" Concentration Measure as a Numbers-Equivalent. *The Review of Economics and Statistics*, **51**, 99–101.

Anderson, M.J., Crist, T.O., Chase, J.M., Vellend, M., Inouye, B.D., Freestone, A.L., Sanders, N.J., Cornell, H.V., Comita, L.S., Davies, K.F., Harrison, S.P., Kraft, N.J.B., Stegen, J.C. & Swenson, N.G. (2011) Navigating the multiple meanings of β diversity : a roadmap for the practicing ecologist. *Ecology Letters*, **14**, 19–28.

Arbia, G. (1989) *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Kluwer, Dordrecht.

Bickenbach, F., Bode, E. & Krieger-Boden, C. (2013) Closing the gap between absolute and relative measures of localization, concentration or specialization. *Papers in Regional Science*, **92**, 465–480.

Chao, A. (2004) Species richness estimation. N. Balakrishnan, C.B. Read & B. Vidakovic, eds., *Encyclopedia of Statistical Sciences*. Wiley, New York, 2nd ed. edition.

Combes, P.P. & Overman, H.G. (2004) The spatial distribution of economic activities in the European Union. J.V. Henderson & J.F. Thisse, eds., *Handbook of Urban and Regional Economics*, volume 4, chapter 64, pp. 2845–2909. Elsevier. North Holland, Amsterdam.

Cousins, S.H. (1991) Species diversity measurement : Choosing the right index. *Trends in Ecology & Evolution*, **6**, 190–192.

Cutrini, E. (2010) Specialization and Concentration from a Twofold Geographical Perspective : Evidence from Europe. *Regional Studies*, **44**, 315–336.

Daróczy, Z. (1970) Generalized information functions. *Information and Control*, **16**, 36–51.

Dauby, G. & Hardy, O.J. (2012) Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species. *Ecography*, **35**, 661–672.

Davis, H.T. (1941) *The theory of econometrics*. The Principia Press, Bloomington, Indiana.

Ellison, G. & Glaeser, E.L. (1997) Geographic Concentration in U.S. Manufacturing Industries : A Dartboard Approach. *Journal of Political Economy*, **105**, 889–927.

Gregorius, H.R. (1991) On the concept of effective number. *Theoretical population biology*, **40**, 269–83.

Gregorius, H.R. (2010) Linking Diversity and Differentiation. *Diversity*, **2**, 370–394.

Havrda, J. & Charvát, F. (1967) Quantification method of classification processes. Concept of structural alpha-entropy. *Kybernetika*, **3**, 30–35.

Hill, M.O. (1973) Diversity and Evenness : A Unifying Notation and Its Consequences. *Ecology*, **54**, 427–432.

Houdebine, M. (1999) Concentration Géographique des Activités et Spécialisation des Départements Français. *Economie et Statistique*, **326-327**, 189–204.

Hurlbert, S.H. (1971) The Nonconcept of Species Diversity : A Critique and Alternative Parameters. *Ecology*, **52**, 577–586.

Jost, L. (2006) Entropy and diversity. *Oikos*, **113**, 363–375.

Jost, L. (2007) Partitioning diversity into independent alpha and beta components. *Ecology*, **88**, 2427–2439.

Keylock, C.J. (2005) Simpson diversity and the Shannon-Wiener index as special cases of a generalized entropy. *Oikos*, **109**, 203–207.

Krugman, P. (1991) *Geography and Trade*. MIT Press, London.

Maasoumi, E. (1993) A compendium to information theory in economics and econometrics. *Econometric Reviews*, **12**, 137–181.

MacArthur, R.H. (1955) Fluctuations of Animal Populations and a Measure of Community Stability. *Ecology*, **36**, 533–536.

MacArthur, R.H. (1965) Patterns of species diversity. *Biological Reviews*, **40**, 510–533.

Marcon, E., Hérault, B., Baraloto, C. & Lang, G. (2012) The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity. *Oikos*, **121**, 516–522.

Marcon, E., Scotti, I., Hérault, B., Rossi, V. & Lang, G. (2014) Generalization of the Partitioning of Shannon Diversity. *Plos One*, **9**, e90289.

Marshall, A. (1890) *Principle of Economics*. Macmillan, London.

May, R.M. (1990) Taxonomy as Destiny. *Nature*, **347**, 129–130.

Mendes, R.S., Evangelista, L.R., Thomaz, S.M., Agostinho, A.A. & Gomes, L.C. (2008) A unified index to measure ecological diversity and species rarity. *Ecography*, **31**, 450–456.

Openshaw, S. & Taylor, P.J. (1979) A million or so correlation coefficients : three experiments on the modifiable areal unit problem. N. Wrigley, ed., *Statistical Applications in the Spatial Sciences*, pp. 127–144. Pion, London.

Patil, G.P. & Taillie, C. (1982) Diversity as a concept and its measurement. *Journal of the American Statistical Association*, **77**, 548–561.

Pavoine, S., Marcon, E. & Ricotta, C. (2016) 'Equivalent numbers' for species, phylogenetic or functional diversity in a nested hierarchy of multiple scales. *Methods in Ecology and Evolution*, **7**, 1152–1163.

Pielou, E.C. (1975) *Ecological Diversity*. Wiley, New York.

Rao, C.R. (1982) Diversity and dissimilarity coefficients : a unified approach. *Theoretical Population Biology*, **21**, 24–43.

Rényi, A. (1961) On Measures of Entropy and Information. J. Neyman, ed., *4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 547–561. University of California Press, Berkeley, USA.

Shannon, C.E. (1948) A Mathematical Theory of Communication. *The Bell System Technical Journal*, **27**, 379–423, 623–656.

Shannon, C.E. & Weaver, W. (1963) *The Mathematical Theory of Communication*. University of Illinois Press.

Simpson, E.H. (1949) Measurement of diversity. *Nature*, **163**, 688.

Theil, H. (1967) *Economics and Information Theory*. Rand McNally and Company, Chicago.

Tilman, D., Knops, J., Wedin, D., Reich, P., Ritchie, M. & Siemann, E. (1997) The Influence of Functional Diversity and Composition on Ecosystem Processes. *Science*, **277**, 1300–1302.

Tsallis, C. (1988) Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, **52**, 479–487.

Tsallis, C. (1994) What are the numbers that experiments provide? *Química Nova*, **17**, 468–471.

Ulanowicz, R.E. (2001) Information theory in ecology. *Computers & Chemistry*, **25**, 393–399.

Webb, C.O., Losos, J.B. & Agrawal, A.A. (2006) Integrating Phylogenies into Community Ecology. *Ecology*, **87**, S1–S2.

Weber, A. (1909) *Über den Standort der Industrien*. Tübingen. English translation edited in 1971, "Theory of the location of industries", Russell & Russell.

Whittaker, R.H. (1960) Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, **30**, 279–338.

Wilson, E.O. & Peter, F.M., eds. (1988) *Biodiversity*. The National Academies Press, Washington, D.C.

Wright, S. (1931) Evolution in Mendelian Populations. *Genetics*, **16**, 97–159.