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

Eric Marcon[1*]

Résumé
Les mesures de la concentration spatiale et de la spécialisation en économie sont très similaires à celles de la biodiversité et de la valence des espèces en écologie. Les développements méthodologiques sont plus avancés en écologie, ce qui motive ce travail de transfert interdisciplinaire. L'entropie est la notion fondamentale, issue de la physique statistique et la théorie de l'information, utilisée pour mesurer la concentration et la spécialisation. La notion de nombre effectif, qui est un nombre de catégories dans une distribution idéale simplifiée, est introduite. La décomposition de la diversité totale d'une distribution (la localisation globale en économie) en concentration ou spécialisation absolue et relative et en réplication, est présentée. L'ensemble fournit un cadre théorique complet et robuste pour mesurer la structuration spatiale en espace discret.

Keywords
entropie, nombres de Hill, diversité, valence, concentration, spécialisation

[1]UMR EcoFoG, AgroParistech, CNRS, Cirad, INRA, Université des Antilles, Université de Guyane.
Campus Agronomique, 97310 Kourou, France.
***Corresponding author** : eric.marcon@agroparistech.fr, https://ericmarcon.github.io/

Table des matières

1. Introduction

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), qui va de pair avec la spécialisation (Houdebine, 1999 ; Cutrini, 2010). De nombreuses mesures de concentration spatiale applicables aux données discrètes (par opposition aux mesures en espace continu traitées par exemple par Marcon and Puech, 2017) ont été développées (pour une revue, voir par exemple Combes and Overman, 2004) mais un cadre méthodologique complet reliant concentration, spécialisation et mesures d'inégalité en général fait défaut, bien qu'il ait été ébauché plusieurs fois (Brülhart and Traeger, 2005 ; Mori et al., 2005 ; Bickenbach and Bode, 2008 ; Cutrini, 2010).

Parallèlement, la mesure de la diversité biologique devenue biodiversité (Wilson and Peter, 1988) et, dans une moindre mesure, de la valence des espèces (Levins, 1968) ont fait l'objet d'une abondante littérature en écologie statistique (Pielou, 1975 ; Patil and Taillie, 1982 ; Magurran, 1988, etc). Elle s'est largement inspirée de la théorie de l'information (Shannon, 1948) et de la physique statistique (Dewar and Porté, 2008). Les mesures de diversité fondées sur l'entropie constituent l'état de l'art en écologie (Marcon, 2017).

En économie, Theil (1967) a proposé des mesures d'inégalité et de concentration spatiale similaires à l'entropie de Shannon, mais les avancées méthodologiques ultérieures sont restées en retrait de celles de la mesure de la biodiversité. L'objectif de cet article est de transférer à la discipline de l'économie géographique les derniers développements de la mesure de la biodiversité pour compléter ses définitions de la concentration spatiale et de la spécialisation. Les emprunts très nombreux de méthodes entre disciplines éloignées seront soulignés.

L'entropie et ses propriétés seront présentées dans la section suivante. L'application de ces méthodes à la mesure de la concentration spatiale et de la spécialisation suivront, avant une dernière section de synthèse consacrée à la diversité jointe, cadre permettant la décomposition complète des mesures de localisation.

2. Méthodes

2.1 Questions similaires et notions opposées

Les méthodes présentées ici ont été développées par la littérature sur la biodiversité. Les écologues ont besoin de mesurer la *diversité* d'une communauté d'êtres vivants, composée de plusieurs espèces dont les effectifs sont connus. Une question moins traitée concerne la *valence* des espèces, c'est-à-dire pour une espèce donnée la diversité des environnements dans lesquels elle est capable de s'installer.

Cette notion a été formalisée sous le nom de largeur de niche par Levins (1968), au sens où la niche écologique est l'ensemble des conditions nécessaires au développement et à la reproduction d'un être vivant. Pour fixer les idées et sans perte de généralité, les exemples traités ici concerneront des arbres dans une forêt. Chaque arbre appartient à une et une seule espèce, et le nombre d'individus de chaque espèce est connu. Les espèces sont situées dans une taxonomie : elles sont regroupées par genres et les genres par familles. Enfin, la forêt est divisée géographiquement en parcelles, elles-mêmes en sous-parcelles. On se limitera ici aux mesures de diversité les plus simples, ne prenant pas en compte les différences plus ou moins grandes entre espèces, taxonomiques par exemple.

En économie géographique, la question probablement la plus traitée est celle de la *concentration spatiale* (Ottaviano and Puga, 1998 ; Combes and Gobillon, 2015), source d'externalités positives (Baldwin and Martin, 2004). Elle est très semblable à la valence des espèces des écologues, mais opposée : une forte concentration est synonyme d'une faible valence.[1] La *spécialisation* (Amiti, 1997) est de même la notion inverse de la diversité. Les exemples traités ici en économie concerneront les établissements industriels des pays d'Europe fournis par la base EuroStat en accès libre. Les établissements ont un nombre d'employés, qui permet leur pondération. Ils appartiennent à un secteur d'activité, ici selon la nomenclature NUTS, qui est une taxonomie similaire à celle des espèces biologiques, et leur localisation par pays peut être détaillée par régions (selon la nomenclature NACE) et leurs subdivisions.

La spécialisation et la concentration spatiale (Cutrini, 2010), comme la diversité et la valence (Gregorius, 2010) sont mathématiquement liées : l'existence de secteurs très concentrés implique celle de régions spécialisées dans ce secteur. Une approche synthétique peut être développée : Cutrini (2010) définit la "localisation globale" à cet effet, qui sera généralisée.

2.2 Données et notations

Les données ont été choisies pour leur accessibilité et leur simplicité : il s'agit ici de présenter des méthodes plus que de traiter en détail des questions économiques complexes. Les applications s'appuieront sur les nombres d'employés par secteur industriel dans 25 pays européens en 2015. Les données sont disponibles en ligne sur la base EuroStat [2], dans le fichier *SBS data by NUTS 2 regions and NACE Rev. 2.*

La nomenclature des secteurs économiques est la NACE (Nomenclature statistique des Activités économiques dans les Communautés Européennes) dans sa révision 2. Seuls les secteurs industriels (code NACE : C) ont été retenus. Les secteurs C12 (manufacture de produits du tabac), C19 (manufacture de coke et produits du pétrole raffiné), C21 (Manufacture de produits pharmaceutiques de base et préparations pharmaceutiques) et C30 (Manufacture d'autres équipements de transport) ont été retirés parce qu'ils présentaient des données manquantes dans des pays majeurs (par exemple, C30 en Belgique).

Parmi les 30 pays disponibles, Chypre, Malte, l'Irlande, le Luxembourg et la Slovénie ont été retirés parce qu'ils comportaient trop de données manquantes. La sélection des données se résume donc à un compromis pour conserver l'essentiel de l'information, tout à fait discutable mais suffisant pour les besoins de démonstration méthodologique de cet article.

Après filtrage, les données se présentent donc sous la forme d'une table (appelée tableau de contingence) dont les 19 lignes sont les secteurs industriels et les 25 colonnes les pays retenus. Chaque cellule du tableau contient le nombre d'employés dans le secteur et le pays considéré, sans données manquantes.

Les secteurs sont indicés par la lettre s et les pays par la lettre i. Les effectifs par secteur et pays sont notés $n_{s,i}$. Les valeurs marginales sont notées n_i (l'effectif du pays i, tous secteurs confondus) et n_s (celui du secteur s, tous pays confondus). Pour alléger l'écriture, le niveau d'agrégation correspondant à l'ensemble des secteurs sera appelé "l'industrie" et celui correspondant à l'ensemble des pays "l'Europe" : n_s sera donc appelé le nombre d'employés travaillant dans le secteur s en Europe. L'effectif total est $n = \sum_s n_s = \sum_i n_i$, égal à 27 419 407. Les tailles relatives des pays et des secteurs sont représentées en annexe. La probabilité qu'une personne choisie au hasard travaille dans le secteur s et le pays i est notée $p_{s,i}$ et estimée par sa fréquence observée $p_{s,i} = n_{s,i}/n$ (pour alléger la notation, la fréquence empirique est notée comme la probabilité théorique plutôt que $\hat{p}_{s,i}$). Les probabilités marginales sont notées p_s et p_i ; elles sont estimées respectivement par n_s/n et n_i/n. Enfin, les probabilités seront aussi considérées par secteur ou par région : $p_{s|i} = p_{s,i}/p_i$ est la probabilité pour une personne du pays i de travailler dans le secteur s. La somme de ces probabilités vaut 1 pour chaque secteur ou chaque région : $\sum_s p_{s|i} = \sum_i p_{i|s} = 1$.

Le vecteur des probabilités $p_{s|i}$ de tous les secteurs dans le pays i est noté $\mathbf{p_{s|i}}$. De même, $p_{i|s}$ est la probabilité, dans le secteur s choisi, qu'une personne travaille dans le pays i et $\mathbf{p_{i|s}}$ est le vecteur des probabilités des pays pour le secteur s. La matrice des probabilités dont les éléments sont $p_{s,i}$ est notée $\mathbf{P}$.

[1] Une espèce de faible valence est dite *spécialiste* en écologie mais ce vocabulaire n'est pas utilisé ici pour éviter toute confusion avec la spécialisation régionale.

[2] http://ec.europa.eu/eurostat/web/regions/data/database

Les données et le code R (R Core Team, 2018) nécessaires pour reproduire l'intégralité des résultats se trouvent en annexe. Le code utilise largement le package *entropart* (Marcon and Hérault, 2015b) consacré à la mesure de la biodiversité.

2.3 L'entropie comme mesure d'incertitude

Les notions étant établies, il s'agit maintenant de les traduire en mesures opérationnelles permettant de comparer la diversité de différentes communautés biologiques (comme les arbres d'une forêt, sans que l'exemple soit limitatif) ou la spécialisation de régions industrielles, de donner un sens concret, facilement compréhensible, à ces mesures, et de caractériser leurs propriétés pour pouvoir les utiliser par exemple dans le cadre de modèles.

La diversité biologique est un déterminant important du fonctionnement des écosystèmes (Chapin et al., 2000). Parmi de très nombreuses mesures développées selon les besoins (Peet, 1974), l'intérêt de l'entropie de Shannon (1948) a été argumenté notamment par Pielou (1975) dans un ouvrage de référence. En économétrie, les travaux de Davis (1941) et surtout Theil (1967) ont ouvert la voie. Le très connu indice de Theil est la différence entre l'entropie de Shannon et sa valeur maximale possible, ce qui illustre l'opposition des approches présentée plus haut en même temps que la convergence des méthodes.

L'entropie est, entre autres, une mesure d'incertitude qu'il est temps de formaliser. Définissons une expérience (par exemple l'échantillonnage d'un arbre au hasard dans une forêt) dont l'ensemble des résultats possibles (l'espèce à laquelle il appartient) est connu. Les résultats sont notés r_s où l'indice s prend toutes les valeurs possibles entre 1 et S, le nombre de résultats possibles. La probabilité d'obtenir r_s est p_s, et $\mathbf{p_s} = (p_1, p_2, \ldots, p_S)$ est l'ensemble (mathématiquement, le vecteur) des probabilités d'obtenir chaque résultat. L'obtention du résultat r_s 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 l'espèce r_s est rare (p_s est petite), son tirage est surprenant. La notion d'information, définie par Shannon, est identique à celle de surprise, plus intuitive. 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 $I(1) = 0$ (l'observation d'un résultat certain n'apporte aucune surprise).

L'entropie est définie comme la moyenne de l'information apportée par tous les résultats possibles de l'expérience. Comme chaque résultat à la probabilité p_s d'être réalisée, la moyenne sur tous les résultats possibles est la moyenne pondérée de $I(p_s)$. L'entropie est définie comme

$$H(\mathbf{p_s}) = \sum_s p_s I(p_s).$$

2.3.1 Entropie de Shannon

Shannon a utilisé la fonction d'information $I(p_s) = -\ln p_s$ pour ses propriétés mathématiques. Elle peut être écrite sous la forme $I(p_s) = \ln(1/p_s)$. L'inverse de la probabilité, $1/p_s$, sera appelé *rareté*[3] : une espèce très rare a une probabilité proche de 0.

La fonction d'information utilisée par Shannon est donc le logarithme de la rareté.

Le terme "entropie" avait été introduit par Clausius en 1865 (*Mémoire IX* dans Clausius, 1868) pour sa nouvelle formulation du second principe de la thermodynamique énoncé par Carnot (1824). Son étymologie grecque signifie *transformation* parce que le second principe concerne la variation d'entropie. Boltzmann a caractérisé l'entropie d'un système complexe (un gaz, donc chaque particule peut avoir plusieurs états possibles) en 1877 (Sharp and Matschinsky, 2015). Shannon (1948) a enfin montré que le nombre d'états possibles d'un système est 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. L'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, donnant naissance à la théorie de l'information.

La pertinence de l'entropie comme mesure de diversité en découle directement : un système est d'autant plus divers qu'il peut avoir un grand nombre d'état possibles ou, de manière équivalente, qu'il est difficile de prévoir l'état dans lequel il se trouve, ou encore qu'il a une entropie élevée.

2.3.2 Entropie généralisée

De nombreuses fonctions d'informations alternatives sont envisageables, y compris les plus exotiques comme $I(p_s) = \cos(p_s\pi/2)$ (Gregorius, 2014).

Parmi elles, trois familles de fonctions paramétrisables se sont imposées : l'entropie généralisée de la littérature des inégalités (Shorrocks, 1980), l'entropie de Rényi (1961), très utilisée jusqu'aux années 2000 pour la mesure de la biodiversité et, plus récemment, l'entropie HCDT détaillée ici.

Tsallis (1988) a proposé une entropie généralisée en physique statistique pour des systèmes ne répondant pas aux propriétés nécessaires à la théorie de Boltzmann. Elle avait été définie par Havrda and Charvát (1967) en cybernétique et redécouverte ensuite, 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).

[3] Patil and Taillie (1982) utilisent le terme *rareté* dans le sens d'*information*, mais cette définition n'a pas été reprise dans la littérature ultérieure.

Sa forme mathématique est :

$$ {}^{q}H(\mathbf{p_s}) = \frac{1}{q-1}\left(1 - \sum_{s=1}^{S} p_s^q\right), $$

où q est un paramètre arbitraire. Lorsque $q = 1$, la formule ne s'applique pas mais la limite de ${}^{q}H(\mathbf{p_s})$ quand $q \to 1$ est l'entropie de Shannon, qui est donc retenue comme définition de ${}^{1}H(\mathbf{p_s})$.

Son intérêt apparaît plus clairement en définissant une généralisation de la fonction logarithme, le logarithme déformé d'ordre q (Tsallis, 1994) comme

$$ \ln_q x = \frac{x^{1-q} - 1}{1-q}. $$

Ici encore, $\ln_q x$ tend vers le logarithme naturel quand q tend vers 1. L'entropie HCDT s'écrit alors comme une généralisation de l'entropie de Shannon :

$$ {}^{q}H(\mathbf{p_s}) = \sum_s p_s \ln_q (1/p_s) $$

Le logarithme déformé est une fonction qui, comme son nom l'indique, déforme la fonction logarithme naturel en changeant sa courbure mais en respectant, quel que soit q, $\ln_q 1 = 0$ et la limite $+\infty$ quand $x \to \infty$. Sa valeur à $x = 0$ est négative mais finie pour $q < 1$. Pour $q \geq 1$, ce n'est pas le cas : $\ln_q x \to -\infty$ quand $x \to 0$. En faisant varier le paramètre q autour de 1, la fonction d'information $\ln_q (1/p_s)$ attribue respectivement une plus ou moins grande surprise aux espèces rares (dont la rareté, $1/p_s$, est grande) quand q croît ou décroît.

On dispose à ce stade d'une définition simple et générale : l'entropie (d'ordre q) d'un système est la surprise moyenne apportée par l'observation d'un de ses individus ; la surprise est le logarithme (d'ordre q) de la rareté. Une communauté biologique est d'autant plus diverse qu'elle est surprenante (que son entropie est grande). Une région est d'autant plus spécialisée que son entropie est faible.

Trois valeurs de q sont particulièrement intéressantes :

- $q = 0$: l'entropie est la richesse, c'est-à-dire S, le nombre d'espèces ou de secteurs, moins 1 ;
- $q = 1$: l'entropie est celle de Shannon. En économétrie, $S - {}^{1}H$ est l'indice de Theil ;
- $q = 2$: l'entropie est l'indice de biodiversité de Simpson (1949), c'est-à-dire la probabilité que deux individus choisis au hasard appartiennent à une espèce différente. En économétrie, son complément à 1, c'est à dire la probabilité que deux individus appartiennent au même secteur, est l'indice de Herfindahl, ou Herfindahl-Hirschman (Hirschman, 1964), qui mesure ici la spécialisation.

Les valeurs négatives de q donnent à une espèce une importance d'autant plus grande qu'elle est rare alors qu'à $q = 0$ toutes les espèces contribuent de façon identique à l'entropie (elles sont simplement comptées, quelle que soit leur probabilité). L'intérêt de ces valeurs est donc limité. Comme leurs propriétés mathématiques sont mauvaises (Marcon et al., 2014), elles ne sont en pratique pas utilisées. Les valeurs de q supérieures à 2 sont peu utilisées parce qu'elles négligent trop les espèces qui ne sont pas les plus fréquentes.

2.3.3 De l'entropie à la diversité

L'entropie a un sens physique : c'est une quantité de surprise ; c'est donc bien plus qu'un indice, qui n'est qu'une valeur arbitraire devant seulement respecter une relation d'ordre pour permettre des comparaisons. Cependant, à l'exception des ordres 0 et 2, la valeur de l'entropie n'a pas d'interprétation intuitive. Les nombres de Hill répondent à ce manque.

Le souhait de Hill (1973) 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'était pas le même selon l'indice de diversité choisi.

Le nombre de Hill d'ordre q est le nombre d'espèces équiprobables donnant la même valeur d'entropie que la distribution observée, autrement dit un *nombre effectif* d'espèces, encore appelé *nombre équivalent*. Le concept a été défini rigoureusement par Gregorius (1991), d'après Wright (1931) qui avait le premier défini la taille effective d'une population en génétique : 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.

Formellement, ils sont simplement l'exponentielle déformée de l'entropie HCDT (Marcon et al., 2014). La fonction exponentielle déformée d'ordre q est la fonction réciproque du logarithme déformé, dont la valeur est

$$ e_q^x = [1 + (1-q)x]^{1/(1-q)}. $$

Le nombre de Hill d'ordre q, appelé simplement *diversité d'ordre q* (Jost, 2006) est donc

$$ {}^{q}D(\mathbf{p_s}) = e_q^{{}^{q}H(\mathbf{p_s})}. $$

La formulation explicite à partir des probabilités est :

$$ {}^{q}D(\mathbf{p_s}) = \left(\sum_s p_s^q\right)^{1/(1-q)}. $$

Ces résultats avaient déjà été obtenus avec une autre approche par MacArthur (1965) et repris par Adelman (1969) dans la littérature économique. Aussi, la mesure d'inégalité `d'@Atkinson1970` est très similaire aux nombres de Hill.

L'utilisation rigoureuse du vocabulaire permet de lever toute ambiguïté (Jost, 2006) : on réservera les

termes diversité, spécialisation, valence et concentration aux nombres de Hill, et ils ne seront pas employés pour l'entropie (qui pourra être appelée *indice* de diversité ou de concentration).

2.3.4 Profils de diversité

La diversité étant exprimée dans la même unité (un nombre d'espèces) quel que soit son ordre, il est possible de tracer un profil de diversité, c'est-à-dire la valeur de ${}^qD$ en fonction de q. Les courbes de deux communautés peuvent se croiser parce que le poids des espèces rares diminue avec l'augmentation de q. Si ce n'est pas le cas, la relation d'ordre entre les communautés est bien définie (Tothmeresz, 1995).

2.4 La décomposition de l'entropie

La notion de diversité β a été introduite par Whittaker (1960) comme le degré de différenciation des communautés biologiques. La question traitée ici est celle de la décomposition de la diversité de données agrégées (la diversité des secteurs économiques en Europe) à un niveau plus détaillé (par pays). La diversité du niveau le plus agrégé a été appelée γ par Whittaker, la diversité moyenne des niveaux détaillés α, et la différenciation entre les niveaux détaillés β. Il est évident que les diversités γ et α sont de même nature : seul le niveau de détail des données diffère. En revanche, la caractérisation de la diversité β a généré des controverses (Ellison, 2010).

En économie, la décomposition des mesures d'inégalité a suivi une voie parallèle à celle des écologues (Bourguignon, 1979). Celle de la concentration spatiale est restée limitée à l'entropie de Theil (Mori et al., 2005 ; Cutrini, 2010) à l'exception notable de Brülhart and Traeger (2005) qui ont utilisé l'entropie généralisée de Shorrocks (1980).

Jost (2007) a montré que la décomposition de l'entropie est additive : l'entropie β est la différence entre les entropies γ et α. Marcon et al. (2012) ont ensuite interprété l'entropie β comme l'information supplémentaire apportée par la connaissance des distributions désagrégées ($\mathbf{p_{s|i}}$ pour chaque pays i) en plus de celle des données agrégées ($\mathbf{p_s}$ pour l'Europe entière), c'est-à-dire une entropie relative. La divergence de Kullback and Leibler (1951) est bien connue des économistes sous le nom d'entropie relative de Theil (Conceição and Ferreira, 2000). La différence entre l'entropie γ d'ordre 1 et la moyenne des entropies d'ordre 1 des distributions désagrégées est la moyenne des divergences de Kullback-Leibler correspondantes (Rao and Nayak, 1985), appelée par les physiciens statistiques "divergence de Jensen-Shannon". Marcon et al. (2014) ont généralisé ce résultat à tous les ordres de l'entropie HCDT : l'entropie β est la moyenne sur tous les pays de la divergence de Kullback-Leibler généralisée entre les distributions $\mathbf{p_{s|i}}$ et $\mathbf{p_s}$, elle-même définie comme la moyenne sur tous les secteurs du gain d'information apporté par la connaissance de la distribution désagrégée :

$$ {}^q_\beta H(\mathbf{P}) = \sum_i p_i \sum_s p_{s|i} [\ln_q (1/p_{s|i}) - \ln_q (1/p_s)] $$

.

Comme l'entropie γ et α, l'entropie β peut être transformée en un nombre effectif qui est le nombre de communautés de même poids, sans espèce commune, qui auraient la même entropie β que les communautés réelles. La décomposition de la diversité est multiplicative : la diversité γ est le produit des diversités α et β.

La décomposition complète est finalement un produit de nombres effectifs : la diversité de l'assemblage de plusieurs communautés biologiques, appelée diversité γ est un nombre effectif d'espèces ; c'est le produit du nombre effectif d'espèces par communauté (diversité α) par le nombre effectif de communautés (diversité β). Elle sera appliquée dans cet article à l'économie des pays européens : le nombre effectif de secteurs économiques de l'Europe (γ) est le produit du nombre effectif moyen de secteurs par pays (α) par un nombre effectif de pays (β).

De même, la valence d'un secteur économique à un niveau agrégé (l'industrie manufacturière) est un nombre effectif de pays (γ), décomposable en un nombre effectif de pays par secteur à un niveau moins agrégé (α) multiplié par un nombre effectif de secteurs (β).

La décomposition sera limitée ici à un seul niveau de désagrégation des données. Elle peut être répétée : les pays peuvent être découpés en régions, les régions en départements… Le nombre effectif de secteurs économiques de l'Europe (γ) peut alors être décomposé en un nombre effectif de pays (β_1) fois un nombre effectif de régions (β_2) fois un nombre effectif de départements (β_3) fois un nombre effectif de secteurs par département (α). La décomposition hiérarchique de la diversité a été traitée notamment par Marcon et al. (2012) ; Richard-Hansen et al. (2015) ; Pavoine et al. (2016).

2.5 Diversité jointe : information mutuelle et réplication

Nous avons vu que l'entropie pouvait être utilisée selon les deux points de vue de la diversité et de la valence (de façon équivalente : la spécialisation et la concentration spatiale). Les données sont les mêmes et peuvent être représentés dans le tableau de contingence dont les lignes représentent par exemple les secteurs industriels alors que les colonnes représentent les pays, chaque cellule du tableau fournissant l'abondance (en nombre d'établissements ou d'employés) d'un secteur dans un pays.

La diversité des pays est calculée en traitant chaque colonne du tableau, la valence des secteurs en traitant chaque ligne. La diversité ${}^qD(\mathbf{p_s})$ de l'Europe entière (définie comme l'agrégation des pays) est obtenue, comme la valence des secteurs agrégés ${}^qD(\mathbf{p_i})$, à partir des probabilités marginales. La diversité ${}^qD(\mathbf{p_{s,i}})$ de l'ensemble des données, tous secteurs et pays confondus, a un grand intérêt, notamment théorique pour

l'entropie de Shannon (Faddeev, 1956 ; Baez et al., 2011) : elle est appelée diversité jointe (Gregorius, 2010).

La différence entre l'entropie jointe et la somme des entropies marginales (celle de l'ensemble des secteurs et celle de l'ensemble des pays), ${}^qH(\mathbf{p_{s,i}}) - {}^qH(\mathbf{p_s}) - {}^qH(\mathbf{p_i})$, s'appelle l'information mutuelle. L'entropie de Shannon (mais pas l'entropie HCDT d'ordre différent de 1) de deux systèmes indépendants s'additionne : si l'appartenance aux pays est indépendante de l'appartenance aux secteurs, c'est-à-dire si la probabilité $p_{s,i}$ est simplement le produit des probabilités p_s et p_i, alors l'information mutuelle de Shannon est nulle. En d'autres termes, l'information mutuelle est l'entropie supplémentaire apportée par la non indépendance des lignes et des colonnes du tableau. Elle est égale aux deux entropies β, celle de la diversité et celle de la valence. Ces propriétés ne sont valables que pour l'entropie de Shannon. Elles ont été utilisées sous différentes formes dans la littérature (par exemple Cutrini, 2009 ; Chao et al., 2013 ; Haedo and Mouchart, 2017).

Quel que soit l'ordre considéré, Gregorius (2010) a montré que la diversité jointe apporte une information supplémentaire importante sur la distribution des abondances qui n'est pas prise en compte par la décomposition de la diversité déjà présentée. L'exemple de la biodiversité est utilisé ici pour simplifier l'exposé. La diversité α est le nombre d'espèces équiprobables dans une communauté type. La diversité β est le nombre de ces communautés types, équiprobables et sans espèce commune. La diversité γ est le produit des deux précédentes, un nombre d'espèces équiprobables résultant de l'assemblage des communautés. Chaque espèce n'apparaît que dans une communauté dans cette représentation. La réplication à l'identique des communautés ne modifie pas les diversités α, β et γ, c'est même une propriété demandée aux mesures de diversité (Hill, 1973). En revanche, la diversité jointe est multipliée par le nombre de réplications (Marcon, 2017) : le rapport entre la diversité jointe et la diversité β mesure la réplication sous la forme d'un nombre effectif, le nombre de répétitions des communautés.

La réplication n'a que peu d'applications pratiques en écologie parce que les données disponibles sont en général des échantillons des communautés étudiées. Leur réplication reflète l'effort d'échantillonnage, qui est un choix de l'expérimentateur. Lorsque les données sont exhaustives ou, plus généralement, lorsque les probabilités marginales des communautés sont interprétables comme leurs tailles, la réplication est une information aussi importante que la diversité.

3. Concentration spatiale et spécialisation

Les méthodes présentées jusqu'ici, issues de la physique et de l'écologie statistique, ont des applications intéressantes en économie. Deux questions seront traitées : celle de la mesure de la concentration spatiale des activités économiques et celle de la décomposition de la diversité jointe.

La concentration spatiale des activités économiques est un sujet important de la littérature (Combes and Gobillon, 2015). La première étape de la compréhension des phénomènes économiques en jeu est la caractérisation de la concentration. Une étape majeure a été franchie par Ellison and Glaeser (1997) qui ont posé clairement le principe d'une mesure relative (la distribution géographique d'un secteur industriel est comparée à celle de la taille des régions où elle est considérée, principe résumé sous le titre d'approche du jeu de fléchettes) et celui du test statistique de la distribution observée contre sa valeur sous une hypothèse nulle appropriée : une distribution uniforme et indépendante. Ces caractéristiques font défaut à des indices de concentration antérieurs, comme l'indice de Gini (Gini, 1912 ; Ceriani and Verme, 2012) dont la valeur observée ne peut être comparée qu'à ses extrêmes possibles.

La statistique centrale de l'indice d'Ellison et Glaeser pour le secteur s est, avec nos notations, $G_s = \sum_i (p_{i|s} - p_i)^2$, c'est-à-dire la somme des carrés des écarts entre la part du pays i dans l'effectif total du secteur s et la part du pays i dans l'industrie, tous secteurs confondus. En termes mathématiques, G est la distance L^2 entre la distribution observée du secteur s et sa distribution attendue (Haedo and Mouchart, 2017), celle de l'industrie en général.

L'indice relatif de Theil (1967) est parfois utilisé dans le même objectif (Cutrini, 2009) : il mesure aussi l'écart entre la distribution observée et la distribution attendue, mais avec une autre métrique : la divergence de Kullback-Leibler.

L'entropie HCDT permet d'unifier et étendre ces approches. La concentration, absolue puis relative, sera envisagée en premier. La spécialisation suivra.

3.1 Concentration spatiale

3.1.1 Valence et concentration absolue

La valence du secteur s, ${}^qD(\mathbf{p_{i|s}})$ est le nombre effectif de pays qu'il occupe. La valence peut être calculée pour n'importe quel niveau de regroupement sectoriel, ici pour l'industrie entière (code NACE C) ou par secteur détaillé. Un profil de valence peut être tracé pour chaque secteur. Aux faibles ordres, la valence donne une grande importance aux pays peu occupés. À $q = 0$, la valence est simplement le nombre de pays dans lesquels le secteur est présent. Aux grands ordres, seuls les pays occupés majoritairement contribuent à la valence.

La figure 1 présente les profils de valence de l'industrie entière et des secteurs C10 (Manufacture de produits alimentaires) et C20 (Manufacture de produits chimiques) qui s'écartent le plus, parmi tous les secteurs étudiés, de l'industrie entière. La valence est mesurée en nombre effectifs de pays. Tous les secteurs sont présents dans tous les pays (les données sont très agrégées) donc la valence d'ordre 0 est toujours égale au maximum possible, 25. À l'ordre 2, à l'autre extrémité des courbes, 9.3 pays occupés par le même nombre d'employés suffiraient pour obtenir le même niveau de valence que celui observé pour l'ensemble

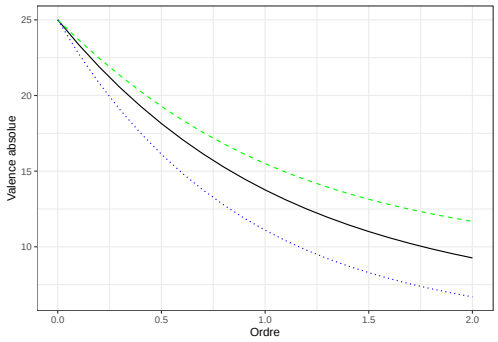

Fig. 1. Profils de valence absolue de l'industrie (courbe pleine, noire), du secteur C10 (pointillés longs, verts) et du secteur C20 (pointillés courts, bleus)

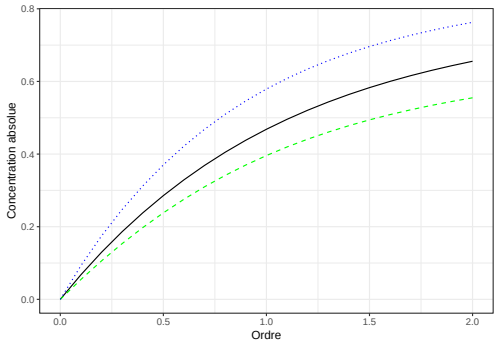

Fig. 2. Profils de concentration absolue de l'industrie (courbe pleine, noire), du secteur C10 (pointillés longs, verts) et du secteur C20 (pointillés courts, bleus).

de l'industrie.

La valence est la notion opposée à celle de concentration. Une transformation simple des valeurs de valence permet de les traduire en niveau de concentration plus conformes à la culture économique. Le complément de la valence au nombre de pays est une bonne mesure de la concentration en tant que nombre effectif de pays délaissés par le secteur étudiés. Il peut être normalisé par le nombre de pays moins 1 pour obtenir une valeur entre 0 et 1 présentée en figure 2.

La valeur de la concentration est la proportion des pays délaissés (en nombres effectifs). Une valeur de 0 signifie que tous les pays sont occupés, une valeur de 1 que tout le secteur est concentré dans un seul pays. L'industrie chimique, C20, est beaucoup plus concentrée que l'industrie en général, alors que l'industrie agro-alimentaire, C10, l'est nettement moins. Ces résultats sont valides à tous les ordres, sauf près de 0, lorsque la présence seule du secteur compte, quelle que soit son abondance.

Cette mesure de la valence ou de la concentration est absolue (Brülhart and Traeger, 2005) : elle ne compare le nombre effectif de secteurs à aucune référence externe. Pour son interprétation, une comparaison à une autre mesure absolue (la concentration à un niveau plus agrégé) est nécessaire (Marcon and Puech, 2017).

3.1.2 Concentration relative

La valence absolue a été calculée au niveau des secteurs désagrégés (C10 et C20) et au niveau de l'industrie entière, dont les effectifs ont été obtenus par agrégation de ceux des secteurs. En reprenant la terminologie de la décomposition de la biodiversité, la valence absolue de l'industrie entière est la valence γ, égale au produit de la valence α (la moyenne de celle des secteurs désagrégés) par la valence β, nombre effectif de secteurs équiprobables ne partageant aucun pays.

La décomposition de l'entropie est la même, mais l'entropie γ est la somme des entropies α et β. L'entropie β est, comme on l'a vu, la divergence de Jensen-Shannon généralisée entre la distribution de chaque secteur et la distribution agrégée de l'industrie. Aux ordres particuliers $q = 1$ et $q = 2$, cette divergence est la moyenne, pondérée par les poids des secteurs, de l'entropie relative de Theil et de la statistique G_s d'Ellison et Glaeser. Ces indices classiques de concentration spatiale sont des divergences de Kullback-Leibler généralisées, en d'autres termes des entropies β d'ordres particuliers, donnant une importance différente aux pays à faibles effectifs : l'indice d'Ellison et Glaeser, d'ordre 2, ne prend en compte que les implantations dominantes.

L'entropie β mesure la concentration relative et non la valence relative : les entropies α et β ont des propriétés fondamentalement différentes. Elle intègre une référence (la distribution de l'industrie tous secteurs confondus) et a donc une valeur attendue, 0, si la distribution de l'industrie considérée est identique à celle de référence.

Sa valeur n'est pas interprétable simplement : il faut recourir au nombre effectif de secteurs dont l'interprétation est intuitive.

Dans le cadre de la décomposition présenté plus haut, la concentration relative moyenne est le rapport de la valence γ sur la valence α : elle s'applique à l'ensemble des secteurs mais ne donne pas d'information sur un secteur particulier.

Elle doit donc être détaillée pour chaque secteur : la concentration relative du secteur s est définie comme le rapport entre la valence absolue de l'ensemble de l'industrie (γ) et sa valence absolue propre :

$$ {}^qC_s = {}^qD(\mathbf{p_i})/{}^qD(\mathbf{p_{i|s}}). $$

C'est un nombre effectif de secteurs : si tous les secteurs avaient une valence égale au nombre effectif de pays ${}^qD(\mathbf{p_{i|s}})$, il en faudrait ${}^qC_s$ pour obtenir une industrie dont la valence serait de ${}^qD(\mathbf{p_i})$ pays effectifs.

La valeur de la concentration relative est visible sur la figure 1 : elle est égale au rapport entre les valeurs des profils de valence de l'industrie et du secteur considéré. Pour l'industrie chimique (C20), elle varie de 1 (à l'ordre 0) à 1.4 à l'ordre 2 : 1.4 secteurs effectifs de valence effective celle du secteur C20, soit 6.7 pays,

formeraient une industrie dont la valence serait celle observée pour l'industrie européenne, $6.7 \times 1.4 = 9.3$ pays effectifs.

La concentration est inférieure à 1 pour l'industrie agro-alimentaire (C10) : 0.79 secteurs effectifs de même caractéristiques que le secteur C10 suffiraient pour constituer la valence de l'industrie européenne. En d'autres termes, le secteur C10 est relativement dispersé.

La concentration relative et la valence absolue (figure 1) sont liées : leur produit est la valence absolue de l'ensemble des secteurs, prise comme référence. La concentration absolue (figure 2) va donc de pair avec la concentration relative, mais les informations qu'elles fournissent sont différentes.

Dans la littérature économique, l'entropie relative a été utilisée pour mesurer la concentration relative par Brülhart and Traeger (2005). Mori et al. (2005) ont utilisé la divergence de Kullback-Leibler entre la distribution d'un secteur et celle de la surface (au lieu du nombre d'employés travaillant dans l'industrie) des régions du Japon pour mesurer la concentration topographique (et non relative) des secteurs. Rysman and Greenstein (2005) ont proposé un test de la concentration relative d'un secteur fondé sur le rapport de vraisemblance des distributions du secteur et de l'industrie entière, qui est simplement la divergence de Kullback-Leibler (voir Mori et al., 2005, pour une présentation détaillée des liens entre les deux approches). Alonso-Villar and Del Río (2013) ont proposé une décomposition de l'entropie généralisée mais se sont limités en pratique à l'ordre 1.

L'entropie relative de Theil a été utilisée pour comparer l'évolution de la concentration spatiale dans le temps (par exemple, Cutrini, 2010) puisqu'elle obéit bien à une relation d'ordre comme toute entropie. Enfin, Bickenbach et al. (2013) ont combiné l'indice de Theil (absolu) et l'indice relatif de Theil pour mieux décrire la concentration spatiale en les appliquant à des secteurs économiques différents (industrie et services).

3.2 Spécialisation

La mesure de la spécialisation fonctionne exactement de la même manière que celle de la concentration, en échangeant le rôle des lignes et des colonnes du tableau de contingence.

La figure 3 présente les profils de diversité absolue de l'Italie, l'Allemagne, la France et l'Islande et de l'Europe. La diversité est le nombre effectif de secteurs équiprobables qui fourniraient la même diversité que celle observée. Comme précédemment, le niveau d'agrégation des données est tel que tous les secteurs sont représentés dans tous les pays : la richesse, c'est-à-dire la diversité d'ordre 0 est égale au nombre secteurs. Tous les pays sont moins divers que l'Europe : ils sont donc tous spécialisés à des degrés divers. L'Italie l'est assez peu, l'Allemagne l'est plus que la France et l'Islande est le pays le plus spécialisé d'Europe avec moins de 5 secteurs industriels effectifs à l'ordre 2,

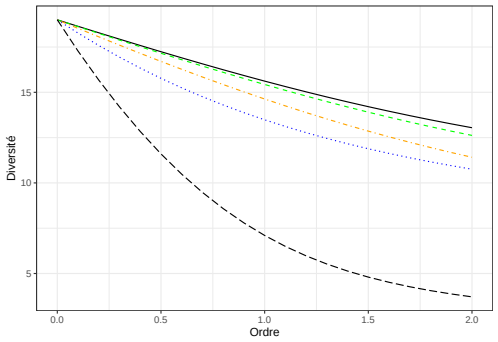

Fig. 3. Profils de diversité de l'Europe (courbe pleine, noire), de l'Italie (pointillés longs, verts), de la France (pointillés alternés, orange), de l'Allemagne (pointillés courts, bleus) et de l'Islande (pointillés très longs, noirs).

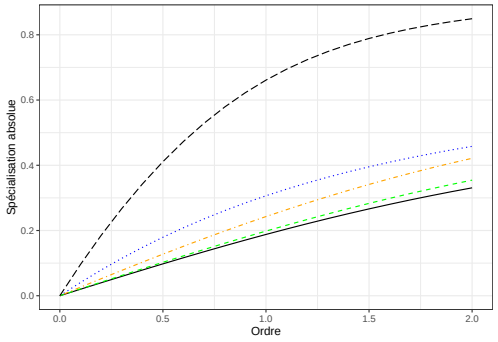

Fig. 4. Profils de spécialisation absolue de l'Europe (courbe pleine, noire), de l'Italie (pointillés longs, verts), de la France (pointillés alternés, orange), de l'Allemagne (pointillés courts, bleus) et de l'Islande (pointillés très longs, noirs).

trois fois moins que l'Italie. L'agro-alimentaire, C10, emploie près de la moitié des employés travaillant dans l'industrie en Islande.

La diversité peut être transformée en spécialisation absolue, comme la valence l'a été en concentration, pour obtenir la figure 4.

Enfin, la spécialisation relative est le rapport entre la diversité absolue de l'Europe entière et celle de chaque pays, visible sur la figure 3. L'Islande est le pays le plus spécialisé relativement, avec une valeur à l'ordre 2 de 3.5 pays effectifs.

3.3 Tests de significativité

Deux approches sont envisageables pour tester les profils de concentration ou spécialisation. Pour fixer les idées et sans perte de généralité, il s'agit ici de tester la spécialisation de l'Italie contre l'hypothèse nulle qu'elle ne serait pas différente de celle de l'Europe entière.

La première formalisation de l'hypothèse nulle est que la distribution des secteurs industriels en Italie est la même que celle de l'Europe entière. Le test porte alors sur la valeur de la divergence de Kullback-Leibler généralisée entre la distribution des secteurs en Italie

et la distribution des secteurs agrégée au niveau européen. C'est l'approche, pour la concentration spatiale, de Mori et al. (2005) à l'ordre 1. La moyenne pour tous les secteurs s de la statistique G_s d'Ellison et Glaeser est égale à la moyenne des divergences d'ordre 2 entre la distribution des secteurs et celle de l'ensemble de l'industrie (Marcon, 2017, section 12.4). L'interprétation de la divergence n'est pas robuste (Jost, 2007) : la statistique testée est l'entropie β mais dans certains cas sa valeur est contrainte par celle de l'entropie α quelle que soit l'entropie γ.

La formalisation alternative est que la spécialisation de l'Italie est égale à celle d'un pays de même taille dont la distribution des secteurs serait celle de l'Europe entière. La statistique testée est un nombre effectif (la diversité absolue ou la spécialisation relative, de façon équivalente). C'est l'approche retenue ici parce qu'elle ne souffre pas du problème de dépendance des entropies α et β.

Le test est réalisé par bootstrap, c'est-à-dire en générant aléatoirement, un grand nombre de fois, de nouvelles données correspondant à l'hypothèse nulle et en calculant la statistique d'intérêt. Les données sont simulées par 1000 tirages dans une loi multinomiale dont les paramètres sont le nombre d'employés travaillant en Italie et les probabilités des secteurs au niveau européen. La spécialisation de chaque simulation est calculée pour les ordres de 0 à 2, par intervalles de 0.1. Les quantiles correspondant à 2,5% et 97,5% des spécialisation simulées constituent les limites de l'enveloppe de confiance de la statistique sous l'hypothèse nulle, à laquelle est comparée la spécialisation réelle.

Le détail du test est présenté en annexe. L'hypothèse nulle n'est pas rejetée à l'ordre 0 : la spécialisation de l'Italie est identique à celle de l'Europe puisque dans les deux cas tous les secteurs sont présents. Dès l'ordre 0.1, l'hypothèse nulle est rejetée : l'Italie est plus spécialisée que l'Europe.

La variabilité de la spécialisation simulée est extrêmement faible parce que les effectifs sont grands et que les employés sont redistribués indépendamment les uns des autres par la loi multinomiale. Pour cette raison, Mori et al. (2005), à partir de données similaires, choisissent de tester la concentration spatiale des établissements en ignorant leurs effectifs. Une bien meilleure hypothèse nulle est que les établissements sont distribués aléatoirement, mais avec leur taille réelle, ce qui augmente fortement l'incertitude sur la spécialisation simulée. Des données individuelles sur les établissements, ou au minimum sur la distribution de leurs tailles, sont nécessaires pour aller plus loin. Avec les données disponibles ici, tous les profils présentés dans les figures 1 à 4 sont significativement différents les uns des autres dès l'ordre 0.1.

4. Diversité jointe

Le tableau de contingence des secteurs et des pays permet de décomposer la diversité totale et d'en tirer plusieurs informations intéressantes. Pour conserver les propriétés de la décomposition, la diversité et la valence ne seront pas transformées en spécialisation et concentration.

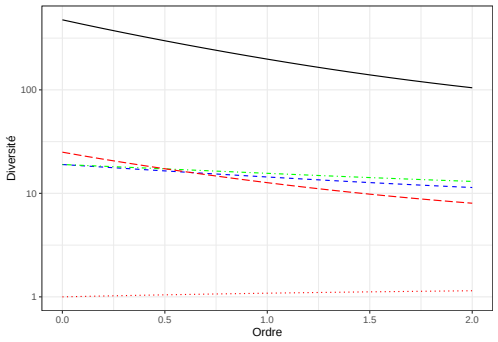

Fig. 5. Profils de diversité des pays européens : diversité jointe (courbe pleine, noire), diversité α intra-pays (pointillés bleus), nombre effectif de pays (diversité β, spécialisation relative moyenne : pointillés courts, rouges), diversité de l'Europe (γ : pointillés longs, verts) et réplication des pays (pointillés longs, rouges). L'échelle de la diversité est logarithmique.

Cutrini (2010) a appelé localisation globale l'information mutuelle du tableau de contingence, c'est-à-dire la concentration relative des secteurs, égale à la spécialisation relative des régions mesurés par la divergence de Kullback-Leibler (l'indice relatif de Theil). Ce n'est qu'une partie de l'information fournie par les données, la diversité ou la valence β, et cette approche n'est valide qu'à l'ordre 1. La diversité jointe exploite toute l'information en combinant diversités ou valences α, β (formant ensemble la diversité γ) et réplication.

4.1 Diversité (spécialisation) des pays

La diversité jointe est décomposée en produit de la diversité α, le nombre effectif de secteurs industriels par pays, de la diversité β, le nombre effectif de pays, et de la réplication, le nombre de répliques de ces pays effectifs. La diversité γ, produit de α et β est celle l'Europe. La décomposition est valide à tous les ordres de diversité, et présentée sous la forme de profils (figure 5).

Les profils sont tracés sur une échelle logarithmique parce que les valeurs sont d'ordres de grandeurs différents et aussi parce que la décomposition multiplicative devient additive sous cette forme : la hauteur de la diversité jointe sur la figure est la somme des hauteurs des diversités α et β et de la réplication.

La diversité de l'Europe (γ, courbe verte) a déjà été présentée en figure 2. Elle est très proche de la diversité α, intra-pays (courbe bleue) : le nombre effectif de pays (β, spécialisation relative, courbe rouge) varie de 1 (à l'ordre 0) à 1.1 à l'ordre 2. Cette valeur est très faible : le maximum possible est le nombre de pays, 25, s'ils ne partagent aucun secteur ; en fait seulement 19 parce que le nombre de secteurs est ici inférieur au

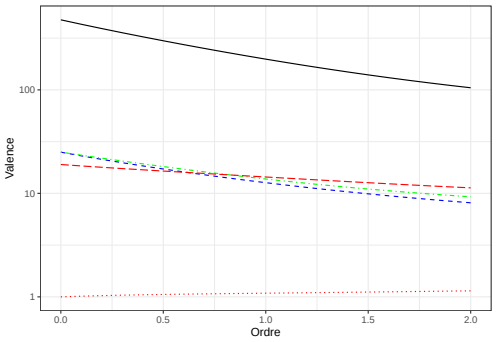

Fig. 6. Profils de valence des secteurs industriels : diversité jointe (courbe pleine, noire), valence α intra-sectorielle (pointillés bleus), nombre effectif de secteurs (concentration relative, valence β : pointillés courts, rouges), valence de l'industrie entière (γ : pointillés longs, verts) et réplication des secteurs (pointillés longs, rouges). L'échelle est logarithmique.

nombre de pays. Les pays présentent donc un certain niveau de spécialisation absolue, mais il est identique à celui de l'Europe entière : leur spécialisation relative est très faible. La spécialisation relative augmente avec l'ordre considéré, c'est-à-dire en négligeant progressivement les secteurs de petite taille : les pays sont un peu plus différents entre eux en ne considérant que les secteurs les plus importants.

La réplication des pays est par conséquent élevée : de 25 à l'ordre 0 (tous les pays abritent tous les secteurs) à 8 à l'ordre 2.

4.2 Valence (concentration) des secteurs

La figure 6 montre la décomposition de la valence des secteurs industriels.

Les résultats sont similaires à ceux de la spécialisation. Les secteurs sont concentrés dans l'absolu mais leur concentration relative est très faible et la réplication est grande.

Cette grande réplication des pays et des secteurs montre qu'à ce niveau d'agrégation des données, l'industrie européenne a une structure peu variable entre pays ou entre secteurs.

5. Conclusion

La théorie de l'information, la physique statistique et l'écologie statistique ont développé des méthodes permettant de définir et mesurer rigoureusement l'incertitude, la diversité, et l'hétérogénéité en général. Les méthodes qui ont été présentées ici permettent d'unifier et d'étendre des approches répandues en économie : la mesure de la concentration spatiale et de la spécialisation par l'entropie, et sa décomposition. Les apports principaux sont un cadre mathématique plus clair, l'utilisation systématique de l'entropie généralisée et la quantification de l'hétérogénéité par des nombres effectifs qui permettent d'interpréter clairement les grandeurs considérées.

Toutes les possibilités méthodologiques n'ont pas été explorées. Un aspect important de la mesure de la biodiversité est son estimation à partir de données échantillonnées plutôt que de données exhaustives (Marcon, 2015), ouvrant la possibilité d'évaluer la concentration ou la spécialisation à partir d'enquêtes plutôt que de bases de données publiques ou commerciales. La diversité fonctionnelle ou phylogénétique (Marcon and Hérault, 2015a) permettrait aussi de prendre en compte la différentiation des secteurs entre eux dans l'évaluation de la spécialisation ou la proximité des régions occupées dans la mesure de la concentration spatiale.

Références

Adelman, M. A. (1969). Comment on the "H" Concentration Measure as a Numbers-Equivalent. *The Review of Economics and Statistics 51*(1), 99–101.

Alonso-Villar, O. and C. Del Río (2013). Concentration of Economic Activity : An Analytical Framework. *Regional Studies 47*(5), 756–772.

Amiti, M. (1997). Specialisation Patterns in Europe. Technical report.

Baez, J. C., T. Fritz, and T. Leinster (2011). A characterization of entropy in terms of information loss. *Entropy 13*(11), 1945–1957.

Baldwin, R. E. and P. Martin (2004). Agglomeration and regional growth. In J. V. Henderson and J.-F. Thisse (Eds.), *Handbook of Urban and Regional Economics*. Amsterdam : Elsevier. North Holland.

Bickenbach, F. and E. Bode (2008). Disproportionality Measures of Concentration, Specialization, and Localization. *International Regional Science Review 31*(4), 359–388.

Bickenbach, F., E. Bode, and C. Krieger-Boden (2013). Closing the gap between absolute and relative measures of localization, concentration or specialization. *Papers in Regional Science 92*(3), 465–480.

Bourguignon, F. (1979). Decomposable Income Inequality Measures. *Econometrica 47*(4), 901–920.

Brülhart, M. and R. Traeger (2005). An Account of Geographic Concentration Patterns in Europe. *Regional Science and Urban Economics 35*(6), 597–624.

Carnot, S. (1824). *Réflexions sur la puissance motrice du feu*. Paris : Bachelier.

Ceriani, L. and P. Verme (2012). The origins of the Gini index : extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *Journal of Economic Inequality 10*(3), 421–443.

Chao, A., Y.-T. Wang, and L. Jost (2013). Entropy and the species accumulation curve : a novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution 4*(11), 1091–1100.

Chapin, F. S. I., E. S. Zavaleta, V. T. Eviner, R. L. Naylor, P. M. Vitousek, H. L. Reynolds, D. U. Hooper, S. Lavorel, O. E. Sala, S. E. Hobbie, M. C. Mack, and S. Díaz (2000). Consequences of changing biodiversity. *Nature 405*(6783), 234–242.

Clausius, R. (1868). *Théorie mécanique de la chaleur*. Paris : Eugène Lacroix.

Combes, P.-P. and L. Gobillon (2015). The empirics of agglomeration economies. In G. Duranton, J. V. Henderson, and W. C. Strange (Eds.), *Handbook of Urban and Regional Economics*, Volume 5, Chapter 5, pp. 247–348. Amsterdam : Elsevier.

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

Conceição, P. and P. Ferreira (2000). The Young Person's Guide to the Theil Index : Suggesting Intuitive Interpretations and Exploring Analytical Applications. Technical report, Austin, Texas.

Cutrini, E. (2009). Using entropy measures to disentangle regional from national localization patterns. *Regional Science and Urban Economics 39*(2), 243–250.

Cutrini, E. (2010). Specialization and Concentration from a Twofold Geographical Perspective : Evidence from Europe. *Regional Studies 44*(3), 315–336.

Daróczy, Z. (1970). Generalized information functions. *Information and Control 16*(1), 36–51.

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

Dewar, R. C. and A. Porté (2008). Statistical mechanics unifies different ecological patterns. *Journal of theoretical biology 251*(3), 389–403.

Ellison, A. M. (2010). Partitioning diversity. *Ecology 91*(7), 1962–1963.

Ellison, G. and E. L. Glaeser (1997). Geographic Concentration in U.S. Manufacturing Industries : A Dartboard Approach. *Journal of Political Economy 105*(5), 889–927.

Faddeev, D. K. (1956). On the concept of entropy of a finite probabilistic scheme. *Uspekhi Mat. Nauk 1*(67), 227–231.

Gini, C. (1912). *Variabilità e mutabilità*. Bologna : C. Cuppini.

Gregorius, H.-R. (1991). On the concept of effective number. *Theoretical population biology 40*(2), 269–83.

Gregorius, H.-R. (2010). Linking Diversity and Differentiation. *Diversity 2*(3), 370–394.

Gregorius, H.-R. (2014). Partitioning of diversity : the "within communities" component. *Web Ecology 14*, 51–60.

Haedo, C. and M. Mouchart (2017). A stochastic independence approach for different measures of concentration and specialization. *Papers in Regional Science in press*.

Havrda, J. and F. Charvát (1967). Quantification method of classification processes. Concept of structural alpha-entropy. *Kybernetika 3*(1), 30–35.

Hill, M. O. (1973). Diversity and Evenness : A Unifying Notation and Its Consequences. *Ecology 54*(2), 427–432.

Hirschman, A. O. (1964). The Paternity of an Index. *The American Economic Review 54*(5), 761–762.

Houdebine, M. (1999). Concentration Géographique des Activités et Spécialisation des Départements Français. *Economie et Statistique 326-327*(6-7), 189–204.

Hurlbert, S. H. (1971). The Nonconcept of Species Diversity : A Critique and Alternative Parameters. *Ecology 52*(4), 577–586.

Jost, L. (2006). Entropy and diversity. *Oikos 113*(2), 363–375.

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology 88*(10), 2427–2439.

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

Kullback, S. and R. A. Leibler (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Levins, R. (1968). *Evolution in Changing Environments : Some Theoretical Explorations*. Princeton University Press.

MacArthur, R. H. (1965). Patterns of species diversity. *Biological Reviews 40*(4), 510–533.

Magurran, A. E. (1988). *Ecological diversity and its measurement*. Princeton, NJ : Princeton University Press.

Marcon, E. (2015). Practical Estimation of Diversity from Abundance Data. *HAL 01212435*(version 2).

Marcon, E. (2017). *Mesures de la Biodiversité*. Kourou, France : UMR EcoFoG.

Marcon, E. and B. Hérault (2015a). Decomposing Phylodiversity. *Methods in Ecology and Evolution 6*(3), 333–339.

Marcon, E. and B. Hérault (2015b). entropart, an R Package to Measure and Partition Diversity. *Journal of Statistical Software 67*(8), 1–26.

Marcon, E., B. Hérault, C. Baraloto, and G. Lang (2012). The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity. *Oikos 121*(4), 516–522.

Marcon, E. and F. Puech (2017). A Typology of Distance-Based Measures of Spatial Concentration. *Regional Science and Urban Economics 62*, 56–67.

Marcon, E., I. Scotti, B. Hérault, V. Rossi, and G. Lang (2014). Generalization of the Partitioning of Shannon Diversity. *Plos One 9*(3), e90289.

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

Mendes, R. S., L. R. Evangelista, S. M. Thomaz, A. A. Agostinho, and L. C. Gomes (2008). A unified index to measure ecological diversity and species rarity. *Ecography 31*(4), 450–456.

Mori, T., K. Nishikimi, and T. E. Smith (2005). A Divergence Statistic for Industrial Localization. *The Review of Economics and Statistics 87*(4), 635–651.

Ottaviano, G. I. P. and D. Puga (1998). Agglomeration in the global economy : A survey of the new economic geography. *The World Economy 21*(6), 707–731.

Patil, G. P. and C. Taillie (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association 77*(379), 548–561.

Pavoine, S., E. Marcon, and C. Ricotta (2016). 'Equivalent numbers' for species, phylogenetic or functional diversity in a nested hierarchy of multiple scales. *Methods in Ecology and Evolution 7*(10), 1152–1163.

Peet, R. K. (1974). The measurement of species diversity. *Annual review of ecology and systematics 5*, 285–307.

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

R Core Team (2018). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing.

Rao, C. R. and T. K. Nayak (1985). Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *IEEE Transactions on Information Theory 31*(5), 589–593.

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

Richard-Hansen, C., G. Jaouen, T. Denis, O. Brunaux, E. Marcon, and S. Guitet (2015). Landscape patterns influence communities of medium- to large-bodied vertebrate in undisturbed terra firme forests of French Guiana. *Journal of Tropical Ecology 31*(5), 423–436.

Rysman, M. and S. Greenstein (2005). Testing for agglomeration and dispersion. *Economics Letters 86*, 405–411.

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

Sharp, K. and F. Matschinsky (2015). Translation of Ludwig Boltzmann's paper "on the relationship between the second fundamental theorem of the mechanical theory of heat and probability calculations regarding the conditions for thermal equilibrium". *Entropy 17*(4), 1971–2009.

Shorrocks, A. F. (1980, apr). The Class of Additively Decomposable Inequality Measures. *Econometrica 48*(3), 613.

Simpson, E. H. (1949). Measurement of diversity. *Nature 163*(4148), 688.

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

Tothmeresz, B. (1995). Comparison of different methods for diversity ordering. *Journal of Vegetation Science 6*(2), 283–290.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics 52*(1), 479–487.

Tsallis, C. (1994). What are the numbers that experiments provide ? *Química Nova 17*(6), 468–471.

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*(3), 279–338.

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

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics 16*(2), 97–159.