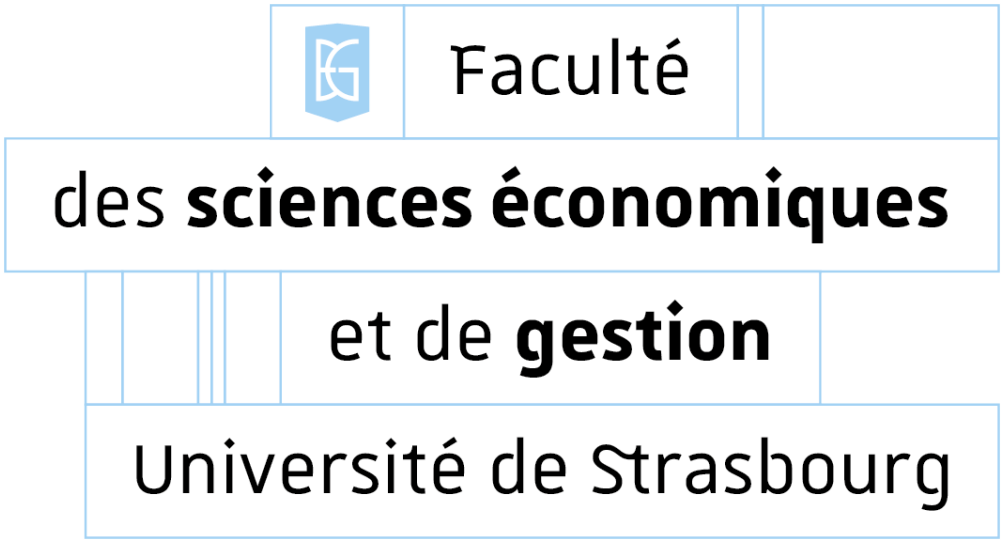
Analyse Scientométrique de la Nouveauté et de la Rupture dans la Recherche et Développement : Impact des Technologies Numériques sur les Objectifs de Développement Durable

Auteur: Valentin Barthel*
Supervisé par: Stefano Bianchini

2 juin 2024



*Magistère Génie Economique 2 et Master 1 Data science pour l'économie et l'entreprise du futur, Faculté des sciences Economiques et de Gestion, Université de Strasbourg, 61 avenue de la Forêt Noire, 67000 Strasbourg, France, valentin.barthel@etu.unistra.fr
Github: https://github.com/valentinb67
Script Python du Mémoire: https://github.com/valentinb67/Master-s-thesis-2-A-scientometric-approach-to-determine-the-effect-of-group-size-on-novelty

Table des matières

1 Introduction

Ce mémoire fait suite à une revue de littérature rédigée en première année de Magistère Génie Economique (2023) : "Comment les outils numériques peuvent-ils contribuer à satisfaire les besoins liés à la gestion énergétique de manière propre et durable ?", dans laquelle nous nous étions intéressé aux avantages liés à l'usage des technologies numériques pour la production et la gestion d'énergie propre et durable. Nous avions développé l'intérêt des technologies numériques pour la contribution en R&D dans le contexte d'une quatrième révolution industrielle (révolution du numérique) et développé le fait que l'on puisse y tirer des externalités pour le développement et le déploiement des technologies vertes. Nous avions discuté de l'intérêt des technologies numériques pour la gestion optimale des énergies avec notamment des méthodes de criblage virtuel pour la découverte de matériaux aux propriétés résistantes, plus isolantes et plus facilement recyclables. Dans le cadre de la production et la gestion de l'énergie, nous avions discuté du potentiel futur déploiement de méthodes novatrices autour de la fusion nucléaire par confinement magnétique assistés par réseaux de neurones ainsi que le déploiement des smart grids. Dans ce mémoire, en adopotant une approche scientométrique, nous allons nous intéresser aux indicateurs de nouveauté afin d'évaluer l'impact de l'utilisation des technologies numériques sur le potentiel de nouveauté des productions scientifiques dans le cadre des Objectifs de Développement Durable (ODD). Cette démarche vise à aller au-delà du comptage traditionnel de citations pour évaluer la qualité d'une production scientifique. Nous allons dans un premier temps expliquer plus en détail l'intérêt des indicateurs de nouveauté dans le cadre de l'Econnomie de l'Innovation avant de relier les approches existantes dans la littérature autour des méthodes visant à évaluer le degré de nouveauté des productions scientifiques ainsi que des points de rupture. Pour finir, nous nous livrerons à une méthode de novelty décrite par Lee et al. (2015) à l'aide de la bibliothèque python "Novelpy" (Pelletier and Wirtz, 2022) dans laquelle nous testerons l'hypothèse de la relation en U inversé entre la Novelty et la taille de l'équipe et discuterons de l'effet de l'utilisation des technologies numériques pour la recherche dans le cadre des ODD à partir d'un échantillon de 1804 articles.

2 L'Intérêt des Indicateurs Mesurant le Degré de Nouveauté (Novelty) et du Degré de Rupture (Disruptivness).

Les scientifiques sont de plus en plus submergés par le volume d'articles publiés. Le nombre total d'articles indexés dans Scopus et Web of Science a augmenté de manière exponentielle ces dernières années ; en 2022, le nombre total d'articles était supérieur de 47% à celui de 2016, ce qui a dépassé la

croissance limitée, voire inexistante, du nombre de scientifiques en exercice (Hanson, Barreiro, Crosetto and Brockington, 2023). Ce résultat montre qu'il devient de plus en plus compliqué pour les auteurs de se distinguer dans la littérature mais également de reconnaître les articles innovants si l'on se réfère à la traditionnelle métrique de citations, en raison des risques de distorsion [1] et de sous-estimation de la véritable contribution scientifique des articles [2] que cette dernière représente. Par ailleurs, le nombre de citation peut-être fortement influencé par la notoriété de l'auteur ou du journal, des pratiques de citation stratégiques et comportements opportunistes.

La dépendance aux métriques de citation pour évaluer la recherche scientifique intensifie le biais de publication, créant un cercle vicieux où les travaux publiés et cités reflètent un aspect biaisé de la recherche scientifique dans laquelle les journaux publient et favorisent d'avantages des résultats positifs que négatifs, des résultats qui seront bien plus cités par les auteurs (Franco, Malhotra and Simonovits (2014), Fortunato, Bergstrom, Börner, Evans, Helbing, Milojević, Petersen, Radicchi, Sinatra, Uzzi et al. (2018)).

Dans les domaines scientifiques, le biais de publication peut mener vers un effet de lock-in en enfermant irréversiblement les auteurs dans un paradigme potentiellement sous-optimal, mettant dans l'ombre d'autres articles qui pourraient, au moins aussi bien, faire évoluer les progrès scientifiques. De ce fait, il s'avère intéressant de déployer des indicateurs de degré de nouveauté et de degré de rupture. Nous verrons que ces indicateurs se montreront pertinents pour détecter des potentielles découvertes clés avec un certain degré d'originalité pour un article scientifique proposant une liaison entre deux disciplines éloignées en citant des combinaisons d'autres articles scientifiques qui n'auraient jamais été liés auparavant (Arts, Melluso and Veugelers, 2023).

2.1 Des Premières Théories en Economie de l'Innovation...

Selon Schumpeter, l'innovation est la création de nouvelles combinaisons de production, par exemple, l'introduction d'un nouveau bien, d'une nouvelle méthode de production, l'ouverture d'un nouveau marché, l'acquisition d'une nouvelle source de matériaux ou de biens semi-finis ou la mise en œuvre d'une nouvelle organisation de toute industrie (Schumpeter and Swedberg). Schumpeter met l'accent sur le caractère cruciale des innovations pour les cycles économiques et de croissance à long terme et introduit la notion de destruction créatrice dans laquelle les anciens concepts et structures sont remplacés par des nouveaux, marquant le passage à un nouveau paradigme à la suite d'une Innovation Radicale (Disruptiv-

1. De nombreuses citations ne signalent pas une influence intellectuelle significative. Une étude mentionnée dans l'article montre que plus de la moitié des articles cités avaient une influence intellectuelle mineure ou très mineure, ce qui signifie que les citations peuvent donner une image faussée de l'importance réelle d'un travail scientifique.

2. La tendance de citer des travaux plus récents et étroitement liés, plutôt que des travaux plus anciens qui ont jeté les bases d'une nouvelle idée scientifique. Cela peut entraîner une sous-estimation de l'impact des articles fondateurs qui ont introduit des concepts innovants mais qui ne sont plus fréquemment cités.

ness).

Pour mesurer l'innovation et son effet, il est nécessaire de développer des indicateurs capables de capturer non seulement des aspects quantitatifs traditionnels (Investissement en R&D, citations, taux de croissances) mais également des indicateurs capable de capturer des aspects qualitatifs de l'innovation et de la recherche en visant à évaluer le potentiel de nouveauté et de radicalité. Ces indicateurs sont essentiels non seulement pour l'évaluation des performances économiques, mais aussi pour comprendre comment les innovations influencent les structures sociales et environnementales.

2.2 ...A la Construction d'Indicateurs

Nous avons vu précédemment que les méthodes de Machine Learning permettait le déploiement des indicateurs mesurant le degré de nouveauté (Novelty) et du degré de rupture (Disruptivness) afin d'aller au-delà du critère de citations. Plusieurs de ces indicateurs seront mis en avant par une multitude d'auteurs, qui proposeront chacun des méthodes avec des avantages et inconvéniants. Nous verrons dans les sections suivantes que ces méthodes s'organisent principalement avec des outils de Machine Learning afin de projeter une distance entre les caractéristiques graphiques pouvant se représenter à l'aide d'un réseau de graph (cf. Figure 3).

2.2.1 Novelty

Conceptuellement, le degré de nouveauté d'une production scientifique reflète sa capacité à introduire de nouvelles idées, des méthodes ou des résultats qui n'avaient jamais été envisagés auparavant, élargissant ainsi la connaissance scientifique. De ce fait, nous pouvons identifier de la nouveauté scientifique en détectant un nouveau vocabulaire au sein d'une littérature, par exemple Shibayama, Yin and Matsumoto (2021) utilisant des méthdoes de vectorisation. Nous retrouverons également parmi ces indicateurs, l'"Atypicality" (Uzzi, Mukherjee, Stringer and Jones, 2013), le "Commonness" (Lee et al., 2015), et plus récemments l'indicateur Novelty de Wang, Veugelers and Stephan (2017) utilisant des méthodes de comparaison de combinaisons de citations entre les articles d'un échantillon.z

2.2.2 Disruptivness

Contrairement au degré de nouveauté, le degré de rupture mesure l'étendue à laquelle une nouvelle idée perturbe le *statu quo* de la littérature scientifique, en d'autres termes, un tel indicateur reflète le caractère potentiel d'une production scientifique à converger vers un nouveau paradigme Wu, Wang and

Evans (2019), Bu, Waltman and Huang (2021). De ce fait, la Disruptivness peut s'apparenter à la destruction créatrice de Schumpeter.

2.3 Le Rôle des Nouvelles Technologies pour le Déploiement des Méthodes d'Evaluation du Degré de Nouveauté et de Rupture

L'évolution des processus d'innovation nous a conduit à une évolution rapide des technologies numériques, qui, en outre, nous permet la créations des indicateurs de nouveauté et de rupture évoqués dans la section précédente. Dans le cadre des algorithmes, nous retrouvons parmi ces technologies, les matrices de cooccurrence (Uzzi et al. (2013), Lee et al. (2015), Foster, Rzhetsky and Evans (2015), Wang et al. (2017)) et des techniques d'embedding textuel dans le cadre des méthodes de Natural Language Processing (NLP) Shibayama et al. (2021) (cf. Figure 4). Les matrices de cooccurrence permettent d'analyser les relations et les interactions entre différents éléments tels que les mots-clés ou les références au sein d'un document. Cette méthode a pour objectif de quantifier la fréquence et la proximité des termes pour déduire des structures de données complexes, permettant d'identifier des combinaisons inhabituelles et distantes d'articles cités, et par conséquent, potentiellement innovantes. Par ailleurs, les méthodes NLP, telles que Word2Vec (Mikolov, Sutskever, Chen, Corrado and Dean, 2013), transforment le texte en vecteurs numériques afin de capturer des aspects sémantiques du langage. D'une part, la vectorisation permet de mesurer la distance conceptuelle entre les termes dans un espace multidimensionnel et d'autre part d'appréhender la profondeur sémantique des interactions documentaires, permettant une perspective importante sur la créativité et l'innovation dans les publications scientifiques.
D'autres part, les avancées en termes de technologies de stockage et de traitement s'avèrent essentielles. En effet, nous avons vu dans la section précédentes que des milliards d'articles sont publiés Hanson et al. (2023). De ce fait, il est nécessaire d'utiliser des espaces de stockage élevées et des Central Processing Unit (CPU) et Graphics Processing Unit (GPU) performents pour réaliser les calculs permettant de réaliser les matrices de cooccurrence et vectorisations. Il est par ailleurs nécessaire de disposer d'une capacité de RAM importante pour manipuler la quantité importante de données mais également de disposer d'un système de stockage pour l'accès rapide et efficace aux données non-structurées.

Nous avons vu dans cette section que les méthodes numériques permettent de pousser les frontières de la compréhension scientifique, permettant aux scientifiques et aux décideurs de mieux évaluer l'originalité (Novelty) et l'impact de la rupture (Disruptivness) aussi bien à l'aide d'avancée scientifiques en Machine Learning qu'en méthodes de stockage et de traitement des données. De ce fait, les notions conceptuelles que représentent les indicateurs de mesure de la nouveauté et de rupture dépendent de l'évolution des

outils computationnels tant dans un cadre du software que de l'hardware. Pour des raisons évidentes de contraintes computationnelles, nous procéderons à une analyse statistique sur l'indicateur Novelpy dans la section suivante.

3 Cas d'étude avec Novelpy : L'Approche de Lee et al., 2015

3.1 Présentation de l'Indicateur "Commonness"

Dans cette partie, nous nous attarderons sur une application de Novelpy, une bibliothèque python développée par Pelletier and Wirtz, 2022 à partir de laquelle nous appliquerons l'approche de degré de nouveauté, développé par Lee et al., 2015. Ces auteurs nous proposent une méthode qui consiste à associer la nouveauté en fonction de la taille de l'équipe, qui, à mesure que la taille de l'équipe augmente, la nouveauté augmentera également jusqu'à un certain point à partir duquel l'effet marginal d'une augmentation de la taille de l'équipe mènera à une nouveauté scientifique moins importante bien que la taille de l'équipe ait une relation continuellement croissance avec la probabilité d'obtenir un article à fort impact. Par ailleurs, la diversité au sein des équipes, que ce soit en termes de domaines scientifiques (interdisciplinarité) ou de tâches, n'a pas d'effet direct sur l'impact, indépendamment de la nouveauté. Cependant, cette diversité contribue de manière significative à la nouveauté des résultats de recherche.

3.1.1 Calcul de la Fréquence des Paires de Journaux

Dans cette sous-section, nous décrivons la méthode utilisée pour mesurer la nouveauté des articles scientifiques. Cette méthode est adaptée de Uzzi et al. (2013) par Lee et al. (2015) et repose sur l'analyse des combinaisons de deux références citées dans les articles. La nouveauté est évaluée en fonction de la rareté des combinaisons de paires de journaux cités.

Pour chacuns des articles présents dans notre base de données, nous extrayons les références citées, toutes les combinaisons par paires de journaux (i, j) sont relevées et enregistrées ainsi que l'année de publication t de chacun des articles de notre base de donnée, permettant d'obtenir l'univers des possibles $U_t(i, j)$ à l'aide de la matrice de cooccurence.

3.1.2 Définition de la Fréquence des Paires de Journaux

L'indicateur "commonness" Lee et al. (2015) compare un réseau observé de journaux cités avec un réseau théorique en se basant sur la fréquence observée versus la fréquence attendue des liens pour une

année donnée (voir Figure 5). Pour chaque paire d'entités i et j, la "commonness" est calculée comme suit :

La fréquence d'une paire de journaux (i, j) en année t est définie par l'équation suivante :

$$\text{Commonness}_{ijt} = \frac{\text{nombre observé de paires } N_{ijt}}{\text{nombre attendu de paires } E_{ijt}}$$

Où :

N_{ijt} est le nombre de paires de journaux i et j observées dans U_t.

E_{ijt} est le nombre attendu de paires de journaux i et j , calculé comme suit :

$$E_{ijt} = \left(\frac{N_{it}}{N_t} \cdot \frac{N_{jt}}{N_t}\right) \cdot N_t$$

Avec :

N_{it} : nombre de paires de journaux incluant le journal i dans U_t.

N_{jt} : nombre de paires de journaux incluant le journal j dans U_t.

N_t : nombre total de paires de journaux dans U_t.

3.1.3 Calcul de la Novelty au Niveau de l'Echantilon des Articles pour une Année ou une Discipline

L'objectif de ce calcul est d'obtenir un score de Lee pour chacun des articles de notre échantillon. Dans un premier temps nous répéterons les procédures de la première partie (récupération des références et de l'année de publication des articles échantillonés). Pour chacun des articles publié en année t, nous enregistrons la $Commonness_{ijt}$ pour chacune de ses paires de journaux citées. Nous trions ces valeurs et enregistrons le 10e percentile comme indication de la fréquence au niveau de l'article afin de réduire le bruit et par conséquent améliorer la fiabilité de la mesure de l'indicateur de Lee. Nous appliquerons ensuite une transformation logarithmique est appliquée pour obtenir une distribution approximativement normale, et un signe négatif est ajouté pour donner la mesure finale de la novelty, la novelty étant l'opposé de la "Commonness". De ce fait, nous avons :

$$\text{Novelty} = -\ln(p_{10}(\text{Commonness}_{ijt}))$$

Contrairement à Uzzi et al. (2013) qui vise à mesurer l'atypicité d'une publication scientifique en comparant le réseau observé des citations à un réseau aléatoire (voir Figure 6), la méthode de Lee et al. (2015) se révèle plus efficace computationnellement. En effet, cette méthode visant à évaluer la fréquence

d'une combinaison de citations par rapport à une fréquence théorique attendue qui se forme sur le degré des journaux cités est une estimation du degré de nouveauté plus simple, exigeant moins de ressources computationnelles au coût d'une précision amoindrie.

3.2 Application

Cette sous-section vise à tester la validité de l'hyppothèse concernant la relation en U inversé entre la Novelty et la taille de l'équipe mais également de tester l'effet de l'utilisation de technologies numériques sur le potentiel de nouveauté dans le cadre de productions scientifiques portant sur les ODD.

3.2.1 Les Données

Nous disposons d'un échantillon de productions scientifiques traitant des ODD, obtenus à partir de la base de données OpenAlex. Cet échantillon constitué de 1804 articles publiés entre 2018 et 2021 contient un ensemble d'informations que nous exploitons tels qu'un identifiant, une liste de référence, les noms des auteurs, le nombre de citations, le champ et sous-champs explorés. A l'aide du package python Novelpy, nous réussirons à obtenir un indicateur Novelty à partir de la Commonness en utilisant la méthode énoncés précédemment. Nous avons réussi à extraire le nombre d'auteurs que nous avons log-transformés (Team Size). Nous avons également log-transformé du nombre de citations (Citations) et à partir de la variable correspondant au sous-champs déduis, si pour l'article observé, celui-ci abordait des thématiques liés aux technologies numériques. Cette distinction se présente par une dummy qui prend la valeur 1 pour les articles discutant des technologies numériques, 0 sinon (Digital subfield).

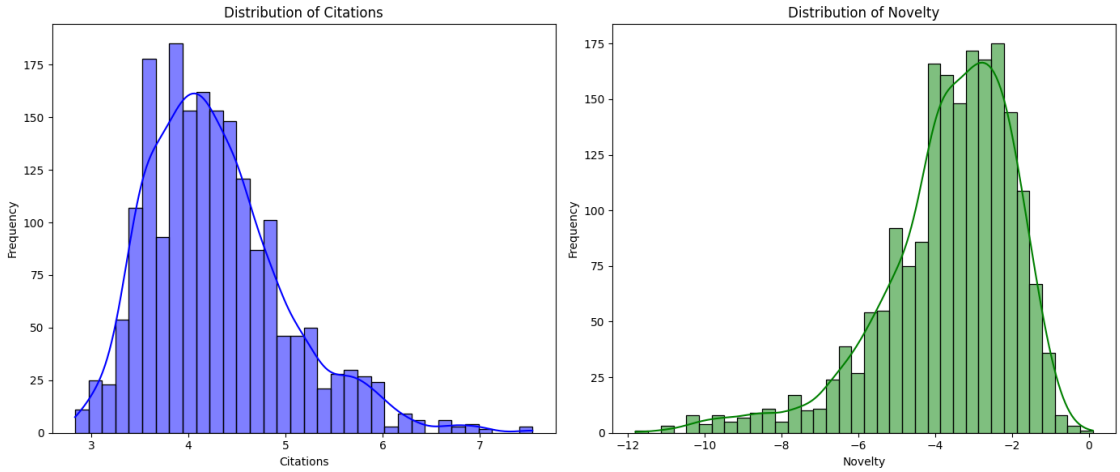


Figure 1. Distribution des Citations et de la Novelty

Il n'est pas étonnant d'observer une distribution centrée sur un petit nombre de citations. Cependant, il est intéressant d'observer qu'un score de Novelty élevé est fréquent. Cette moyenne élevée pourrait s'expliquer par le fait que les ODD sont une thématique large pour lesquels il est fréquent de voir des combinaisons de journaux dédiés à diverses disciplines (sciences juridiques, économiques, biologiques, physiques, ingénierie,...), révélant un potentiel biais dans notre utilisation de l'indicateur. En effet, le fait d'utiliser l'indicateur sur un échantillon d'articles peu spécifique à une discipline mais plutôt large et riche en combinaisons de discipline va limiter la discrimination des productions à potentiel réellement innovants.

Table 1. Correlation Matrix

Variables	**Novelty**	**Team Size**	**Citations**	**Digital subfield**
Novelty	**1.000000**	0.158381	0.228037	0.137155
Team Size	0.158381	**1.000000**	0.283073	-0.082840
Citations	0.228037	0.283073	**1.000000**	-0.019169
Digital subfield	0.137155	-0.082840	-0.019169	**1.000000**

Notes : Team size = log(authors).
Citations = log(num_citations).
Digital subfield is a dummy variable taking values 0 or 1.

Comme attendus, nous pouvons voir qu'il existe une corrélation positive entre la Novelty et la taille de l'équipe (15,8%) [3] ainsi que la Novelty et le fait que la production scientifique observée mentionne une technologie numérique. Il est intéressant de noter que la corrélation entre le nombre de citations et la Novelty n'est que de 23%, justifiant l'intérêt du déploiement d'un indicateur de nouveauté.

3. Lee et al. (2015) avaient obtenu une corrélation de 19%

3.2.2 Résultats

Table 2. OLS Regression Results

	Explanatory variables				
			Novelty		
	(1)	(2)	(3)	(4)	(5)
Intercept	-5.0072^{***}	-5.2233^{***}	-5.0100^{***}	–	–
	(0.223)	(0.222)	(0.276)		
Team size	1.3040^{***}	1.2815^{***}	1.0880^{***}	-3.8099^{***}	-3.7841^{***}
	(0.214)	(0.209)	(0.253)	(0.111)	(0.105)
Team size²	-0.2306^{***}	-0.2147^{***}	-0.1827^{***}	0.8864^{***}	0.8129^{***}
	(0.046)	(0.045)	(0.050)	(0.057)	(0.049)
Digital subfield	–	0.5768^{***}	0.1838	–	-2.5050^{***}
		(0.080)	(0.255)		(0.231)
Team size x Digital subfield	–	–	0.2612^{*}	–	1.9037^{***}
			(0.147)		(0.137)
Observations	1804	1804	1804	1804	1804
R-squared	0.031	0.057	0.058	0.765	0.786
Adj. R-squared	0.030	0.055	0.056	0.765	0.786
F-statistic	23.49	33.19	33.08	3572.00	1955.00

Note :

*** : p < 0.01.

* : p < 0.10.

Team size = log(authors).

Digital subfield is a dummy variable taking values 0 or 1.

Conformément aux résultats de Lee et al. (2015), les modèles (1) à (3) révèlent une relation en U inversé de la taille de l'équipe sur la nouveauté. Conformément à notre hypothèse de départ, le fait que la production scientifique traitant des ODD face référence aux technologies numériques semble accroître le potentiel de nouveauté de la production observé. Par ailleurs, l'association de la taille de l'équipe au fait que l'article fasse référence à une technologie numérique semble avoir un effet positif sur la nouveauté (p < 0.10). Cependant, nous pouvons questionner l'exactitude des coefficients présentés, en effet, ces modèles prévoient un potentiel de nouveauté maximal de 19 auteurs pour les articles qui ne mentionnent pas les technologies numériques, et de 37 auteurs pour ceux qui les mentionnent. Nous avons également réalisé des modèles sans constante afin de supposer qu'une production scientifique ne puisse pas obtenir un score de nouveauté s'il n'a pas d'auteur et qu'il n'existe pas. Ces modèles (4) et (5) montrent des effets inversés par rapport aux trois modèles et par conséquent, ne respectent pas l'hypothèse de l'effet de la taille de

l'équipe en U inversé puisque l'estimation montre l'existence d'un minimum et non d'un maximum ainsi que l'hypothèse suggérant un impact négatif d'un article faisant référence à une technologie numérique sur le potentiel de nouveauté d'une production scientifique traitant des ODD.

L'estimation d'un nombre d'auteur optimal anormalement élevé pour maximiser la Novelty pourrait s'expliquer par un biais de variable omise. Le fait d'omettre une variable corrélée avec la taille de l'équipe pourra biaiser les coefficients et par conséquent, nous conduire à une estimation biaisée d'une nombre optimal d'auteur. Par ailleurs, le fait de retirer la constante dans un contexte où il y a un biais de variable omise conduira à une multicolinéarité (R^2 = 76,5% (4)). Cette multicolinéarité conduira à l'absorption de la variabilité de la Novelty par nos deux variables, et par conséquent, biaisera le signe des coefficients contrairement aux trois modèles dans lesquels la constante absorbe une partie de la variabilité de la Novelty.

4 Discussion

Nous avons exploré le contexte et les méthodes pour évaluer les degrés de nouveauté et de rupture en tant que nouveaux outils de mesure des productions scientifiques. Nous avons vu de l'intérêt du développement de technologies numériques pour le déploiement de ces nouvelles métriques dans le cadre de l'économie de l'innovation. L'indicateur Commonness, bien que computationnellement efficace et permettant de vérifier nos hypothèses sur la relation entre la Novelty et la taille de l'équipe ainsi que l'effet des technologies numériques sur le potentiel de nouveauté des productions scientifiques portant sur les ODD présente des limites importantes. En effet, nous avons vu que l'indicateur Commonness est construit sur des données de combinaisons observées et théoriques à un niveau annuel, de ce fait, cet indicateur se révèle moins efficace pour détecter des tendances ou des schémas de nouveauté s'étendant sur plusieurs années. Ce manque de granularité temporelle peut limiter sa capacité à analyser la dynamique de la nouveauté scientifique, nous contraignant sur un horizon temporel restrictif. L'analyse de la distribution de notre indicateur de Novelty a révélé une moyenne étonnamment élevée de l'indicateur de Novelty. Cette observation, bien que logique dans le cadre de notre échantillon, suggère un biais à notre utilisation de l'indicateur. La moyenne élevée de la Novelty pourrait s'expliquer par le fait que les ODD constituent une thématique large, regroupant des disciplines diverses. Cette diversité entraîne fréquemment des combinaisons de journaux issus de différentes disciplines, augmentant artificiellement la valeur moyenne de la Novelty. En effet, dans un contexte où les domaines de recherche sont variés, les combinaisons de sources diverses deviennent courantes, ce qui peut biaiser l'indicateur en le rendant moins discriminant pour détecter les productions véritablement innovantes. Pour améliorer ce travail, nous proposons plusieurs axes

de développement futurs, nous pourrions utiliser un indicateur "Atypicality" proposé par Uzzi et al. (2013) et offrant une mesure plus détaillé de la Novelty, bien que cela se fasse au prix d'une charge computationnelle plus importante. Au-delà de la contrainte computationnelle, nous pourrions songer à l'application de méthodes de text embedding, comme celles décrites par Shibayama et al. (2021). Cette méthode pourrait enrichir notre compréhension en intégrant des distances sémantiques grâce à la vectorisation afin d'offrir une analyse plus nuancée et sophistiquée de la Novelty. Pour renforcer la précision de nos modèles, il pourrait se révéler intéressant d'intégrer des indicateurs du caractère public ou privé de l'institution d'appartenance de l'auteur, de l'investissement en R&D de l'institution pour l'année "t" considérée, ou encore de son classement par le National Research Council (NRC). Ces variables pourraient fournir un contexte supplémentaire et réduire les biais potentiels dans l'évaluation de la nouveauté.

Annexes

Figures :

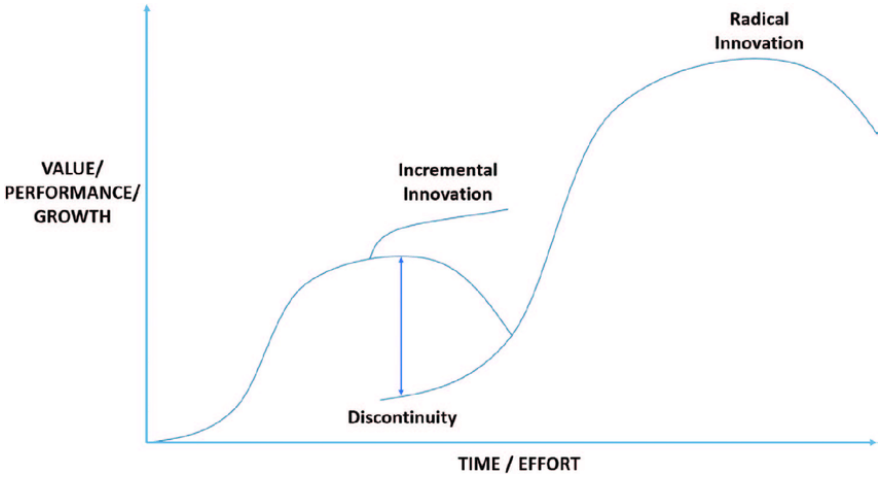


Figure 2. S-Curve of Innovation Salmela (2018)

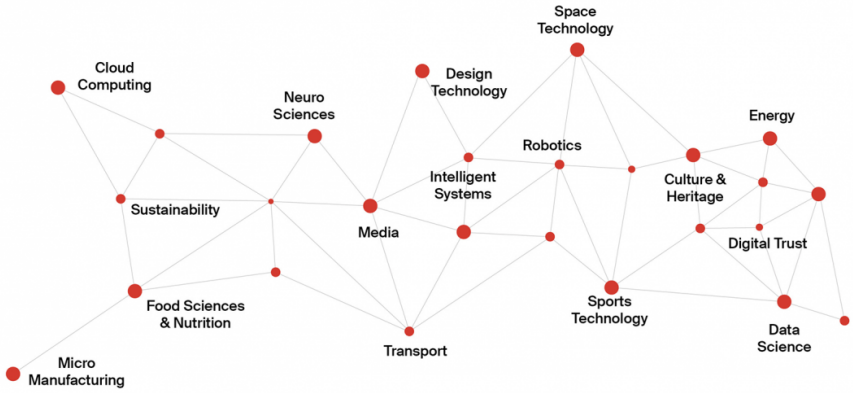


Figure 3. Exemple d'un réseau de graph illustrant une proximité entre disciplines

Source : EPFL

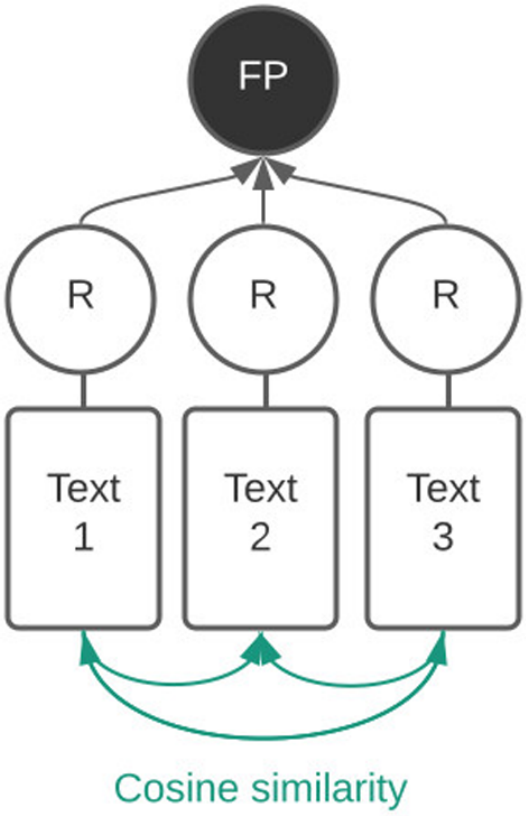


Figure 4. Schema of Shibayama et al. (2021) from Pelletier and Wirtz (2022)

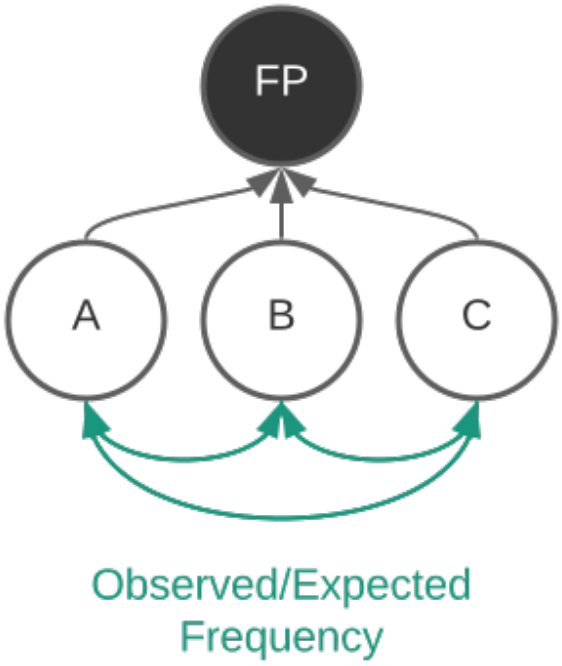
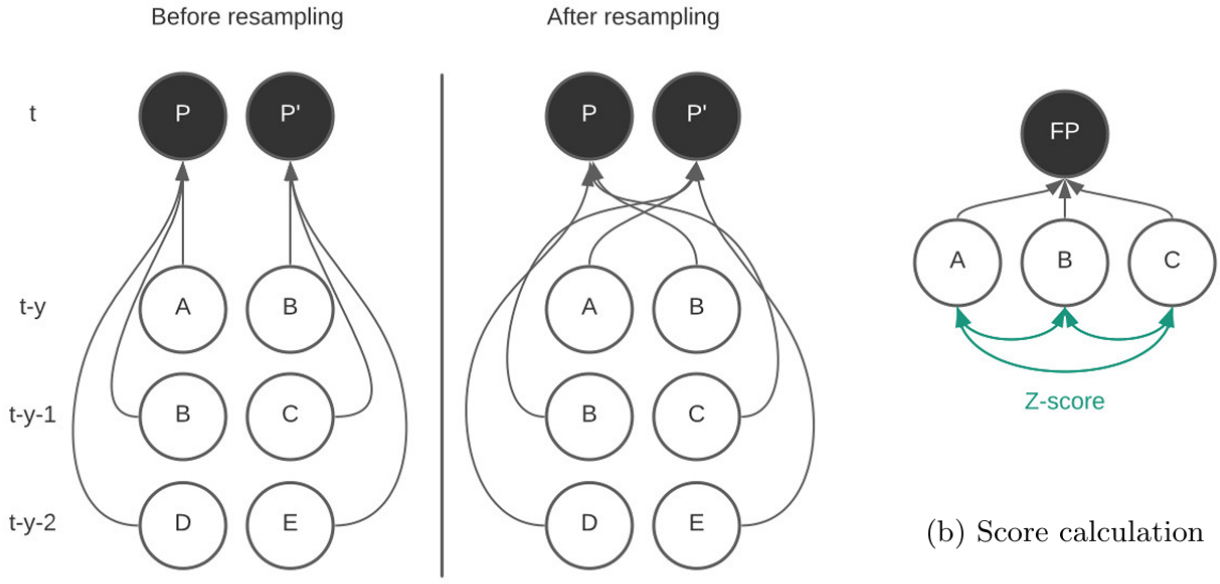


Figure 5. Schema of Lee et al. (2015) from Pelletier and Wirtz (2022)



(a) Resampling strategy

(b) Score calculation

Figure 6. Schema of Uzzi et al. (2013) from Pelletier and Wirtz (2022)

Références

Arts, Sam, Nicola Melluso, and Reinhilde Veugelers, "Beyond Citations : Measuring Novel Scientific Ideas and their Impact in Publication Text," *arXiv e-prints*, 2023, pp. arXiv–2309.

Bu, Yi, Ludo Waltman, and Yong Huang, "A multidimensional framework for characterizing the citation impact of scientific publications," *Quantitative Science Studies*, 2021, *2* (1), 155–183.

Fortunato, Santo, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi et al., "Science of science," *Science*, 2018, *359* (6379), eaao0185.

Foster, Jacob G, Andrey Rzhetsky, and James A Evans, "Tradition and innovation in scientists' research strategies," *American sociological review*, 2015, *80* (5), 875–908.

Franco, Annie, Neil Malhotra, and Gabor Simonovits, "Publication bias in the social sciences : Unlocking the file drawer," *Science*, 2014, *345* (6203), 1502–1505.

Hanson, Mark A, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington, "The strain on scientific publishing," *arXiv preprint arXiv :2309.15884*, 2023.

Lee, You-Na, John P Walsh, and Jian Wang, "Creativity in scientific teams : Unpacking novelty and impact," *Research policy*, 2015, *44* (3), 684–697.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, 2013, *26*.

Pelletier, Pierre and Kevin Wirtz, "Novelpy : A Python package to measure novelty and disruptiveness of bibliometric and patent data," *arXiv preprint arXiv :2211.10346*, 2022.

Salmela, Erno, *Conflicts as Springboard for Metallica's Success* 08 2018.

Schumpeter, Joseph A and Richard Swedberg, *The theory of economic development*, Routledge, 2021.

Shibayama, Sotaro, Deyun Yin, and Kuniko Matsumoto, "Measuring novelty in science with word embedding," *PloS one*, 2021, *16* (7), e0254034.

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones, "Atypical combinations and scientific impact," *Science*, 2013, *342* (6157), 468–472.

Wang, Jian, Reinhilde Veugelers, and Paula Stephan, "Bias against novelty in science : A cautionary tale for users of bibliometric indicators," *Research Policy*, 2017, *46* (8), 1416–1436.

Wu, Lingfei, Dashun Wang, and James A Evans, "Large teams develop and small teams disrupt science and technology," *Nature*, 2019, *566* (7744), 378–382.