Fouille de publications scientifiques pour une analyse bibliométrique de l'activité de recherche sur la déforestation

Julius Akinyemi, Josiane Mothe, Nathalie Neptune

Fouille de publications scientifiques pour une analyse bibliométrique de l'activité de recherche sur la déforestation

Julius Akinyemi*, Josiane Mothe**, Nathalie Neptune**

*Entrepreneur-In-Residence, MIT Media Lab/Founder-CEO UWINCorp Inc.
akinyemi@media.mit.edu
**Institut de Recherche en Informatique de Toulouse
Université de Toulouse
118 Route de Narbonne, F-31062 Toulouse Cedex 9
Josiane.Mothe@irit.fr
Nathalie.Neptune@irit.fr

Résumé. La déforestation est un phénomène très répandu qui touche des portions de territoires assez importantes surtout dans les régions tropicales. La télédétection permet aux chercheurs de suivre et d'analyser l'évolution spatio-temporelle de ce phénomène.

En utilisant la fouille de texte et de méta-données sur les publications scientifiques sur le thème de la déforestation, nous visons à identifier les lieux de la production scientifique sur la déforestation et les collaborations entre chercheurs. L'analyse de ces collaborations nous permet de voir les tendances de la distribution de la production parmi les auteurs, à savoir si elle est concentrée au niveau des auteurs particuliers des pays développés ou bien si elle tend à être répartie de manière équilibrée entre plusieurs pays développés et émergents.

Nous nous appuyons pour cela sur des analyses de réseaux. Par ailleurs, grâce à l'analyse des mots-clés nous identifions les sites touchés par la déforestation auxquels les chercheurs s'intéressent, les forêts tropicales et l'Amazonie, de même que des sujets connexes ayant rapport à l'environnement et à la santé.

1 Introduction

La déforestation est un phénomène environnemental qui peut selon Foley et al. (2005) avoir un impact négatif sur l'écosystème de la terre. Dès 1992, Diegues et al. (1992) passant en revue les liens entre les processus qui conduisent à la déforestation ainsi que ces conséquences dans le basin Amazonien du Brésil, avait estimé que le taux de déforestation était élevé et augmentait rapidement et dangereusement.

La fouille de texte sur les publications scientifiques produites en rapport à la déforestation permet de quantifier les informations sur ces publications ainsi que leur

contenu, Pritchard (1969). Nous avons réalisé une analyse sur ces textes pour voir l'évolution géographique et temporelle de la recherche scientifique sur la déforestation. Cette étude permet d'identifier les principaux acteurs et leur localisation. Par ailleurs, les sujets sur lesquels ils effectuent leurs travaux sont aussi mis en lumière.

Les collaborations entre chercheurs est un aspect important de la recherche. De plus, les travaux de recherche sur la déforestation, de part la nature du phénomène, font souvent appel à des expertises dans des disciplines diverses. Il en découle qu'une analyse des collaborations dans les publications sur la déforestation peut faire ressortir le caractère multidisciplinaire de ces recherches.

Les résultats des analyses effectuées pourront guider des travaux futurs explorant le lien spatio-temporel entre la déforestation et les autres phénomènes liés.

Dans cet article, la méthodologie utilisée pour effectuer les analyses sera d'abord présentée puis les résultats obtenus seront présentés en détails et interprétés. Finalement, des perspectives pour la suite des travaux seront proposées suivies de la conclusion de l'article.

2 Méthodologie

Les données sur les publications ont été collectées à partir du Web of Science Core Collection(`https://webofknowledge.com`). La recherche a été effectuée par sujet en utilisant le terme "deforest*'. Toutes les publications datées de 1975 à 2016 ont été collectées le 23 octobre 2017. Cette approche permet de faire un premier travail de prise de connaissance du domaine. Des requêtes plus sophistiquées auraient pu être utilisées notamment pour retrouver les publications qui ne mentionnent pas explicitement la déforestation tout en y étant liées. Les données ont été analysées avec Tetralogie [1], une plateforme de veille scientifique et technologique qui a été développée à l'Institut de Recherche en Informatique de Toulouse. Pour une description détaillée voir Dousset (2009).

Les fouilles de textes et d'information que nous avons réalisées ont eu pour objectif de répondre aux questions suivantes :

1. Comment le nombre de publications a-t-il évolué avec le temps ?
2. Quels sont les pays au centre de la recherche sur la déforestation ?
3. Quels sont les auteurs qui publient le plus sur le sujet ?
4. Quelle est l'importance et la nature des collaborations entre les chercheurs des différents pays ?
5. À quelles disciplines scientifiques appartiennent les auteurs ?

Des analyses mono et bi-variées ont été réalisées ; les résultats obtenus ont été utilisés pour produire des diagrammes de dispersions, des réseaux de co-auteurs et de catégories de recherche. Finalement, nous avons croisé des données de l'Organisation des Nations Unis pour l'alimentation et l'agriculture et de la Banque Mondiale afin de voir la productivité des pays qui publient le plus par rapport au nombre d'habitants

1. Tetralogie est un logiciel de veille technologique `http://atlas.irit.fr/` utilisée en recherche et en enseignement et pouvant être utilisée à distance contractuellement.

et au produit intérieur brut, pour l'année 2016. Ce croisement met en perspective la production scientifique sur le thème de la déforestation pour un pays par rapport à la taille de sa population et par rapport à la taille de son économie, pour l'année 2016. Ce croisement permet d'avoir une idée de l'effort humain et financier que représente, pour chaque pays leur contribution à la production scientifique sur la déforestation. Les pays des auteurs ont été extraits à partir de l'adresse présente dans les données fournies par le Web of Science. Les différentes orthographes des noms des pays sont prises en compte. Pour certaines publications, l'adresse de l'auteur peut être manquante.

Un total de 16136 publications ont été collectées avec 31772 auteurs dans 149 pays et territoires.

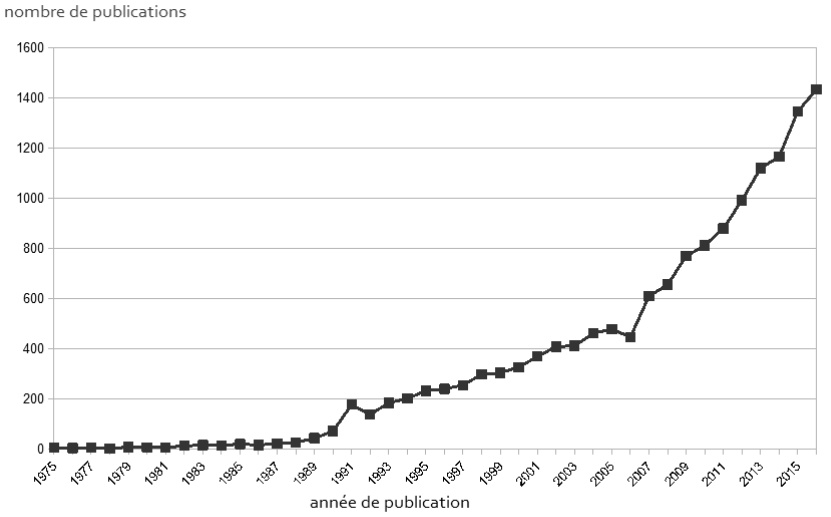

FIG. 1 – *Évolution du nombre de publications avec le temps.*

Le nombre de publications présentes dans la collection est très faible pour les premières années avec moins d'une dizaine de publications par année jusqu'en 1982 qui compte 12 publications. À partir du début des années 90 une augmentation remarquable est constatée dans la production annuelle. Cette tendance à la hausse continue jusqu'en 2016 (voir la figure 1). Cette tendance est similaire à l'évolution constatée sur d'autres domaines notamment les sciences naturelles et les sciences de la santé, selon Bornmann et Mutz (2015). Sur la période des 10 dernières années, de 2006 à 2016, le nombre de publications a augmenté de 220%.

3 Étude relative aux pays

Dans cette section, une analyse des contributions par pays est effectuée puis le réseau de collaboration formé par les pays est présenté. Enfin, le nombre de publications pour chaque pays par rapport à leur produite intérieur brut et à leur population pour l'année 2016 est présenté.

3.1 Contributions par pays et collaborations inter-pays

En regardant les données par pays il en ressort que parmi les dix pays comptant le plus de publications sur la déforestation, sur la période 1996-2016, trois sont des pays émergents et deux sont de l'Amérique latine : le Brésil, la Chine et le Mexique. Voir la figure 2. Il s'agit d'une tendance atypique qui n'est pas constatée sur d'autres domaines. En effet, pour les publications sur les géosciences, uniquement deux pays émergents (Chine et Inde), dont aucun de l'Amérique latine, figurent dans les dix premiers en nombre de publications, voir la figure 3.

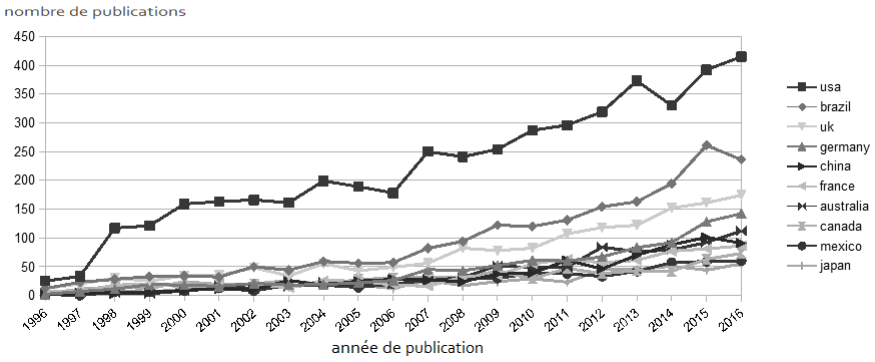

FIG. 2 – *Évolution du nombre de publications par pays avec le temps, pour les dix pays ayant le plus de publications pour toute la période 1996-2016. Les années 1975-1995 ne sont pas représentées compte tenu du faible nombre de publications pour ces années.*

L'évolution de la production de chaque pays peut aussi être évaluée par année par rapport à la production totale du pays. C'est ce que montre la figure 4 pour les 30 dernières années, de 1996 à 2016. Chaque point représente un pourcentage qui est calculé en divisant le nombre de publications du pays pour l'année par la somme des publications du pays pour toutes les années de 1975 à 2016. À partir de 1998, les États-Unis et le Canada cheminent en tête et maintiennent une augmentation de production par rapport à chaque année précédente. Toutefois, cette tendance change à partir de 2008. De 2012 à 2016, les pays ayant le plus augmenté leur production sont l'Allemagne, l'Australie, la Chine et le Brésil.

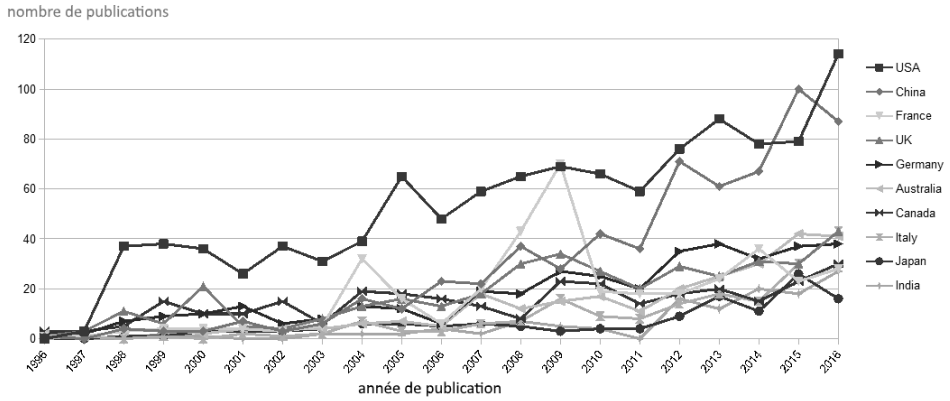

FIG. 3 – *Évolution du nombre de publications contenant le mot clé "geosciences" par pays, pour les dix pays ayant le plus de publications pour toute la période 1996-2016. Un total de* 4641 *publications ont été collectées avec* 13699 *auteurs dans* 87 *pays et territoires.*

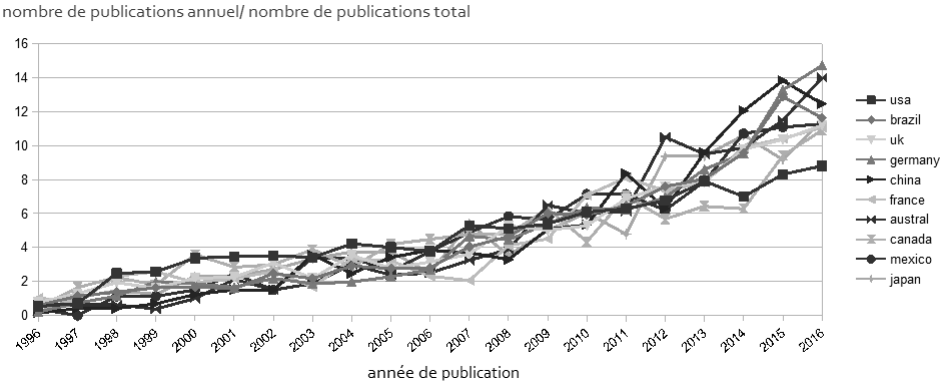

FIG. 4 – *Rapport entre le nombre de publications pour chaque pays sur le nombre total de publications pour le pays, par année, de 1996 à 2016 en pourcentage.*

3.2 Réseau de collaboration entre les pays

Le réseau de collaboration entre les pays fait ressortir un fort niveau de collaboration entre les auteurs issus des différents pays (voir figure 5). Chaque nœud représente un pays et l'intensité des liens entre eux représente le nombre de publications co-écrites par des auteurs de différents pays. On peut observer que tous les pays se retrouvent dans un grand groupe centré principalement autour des États-Unis.

La figure 6 permet de voir les pays ayant le plus de collaborations avec les États-Unis. La taille de chaque nœud représente le nombre de publications du pays. Le Brésil

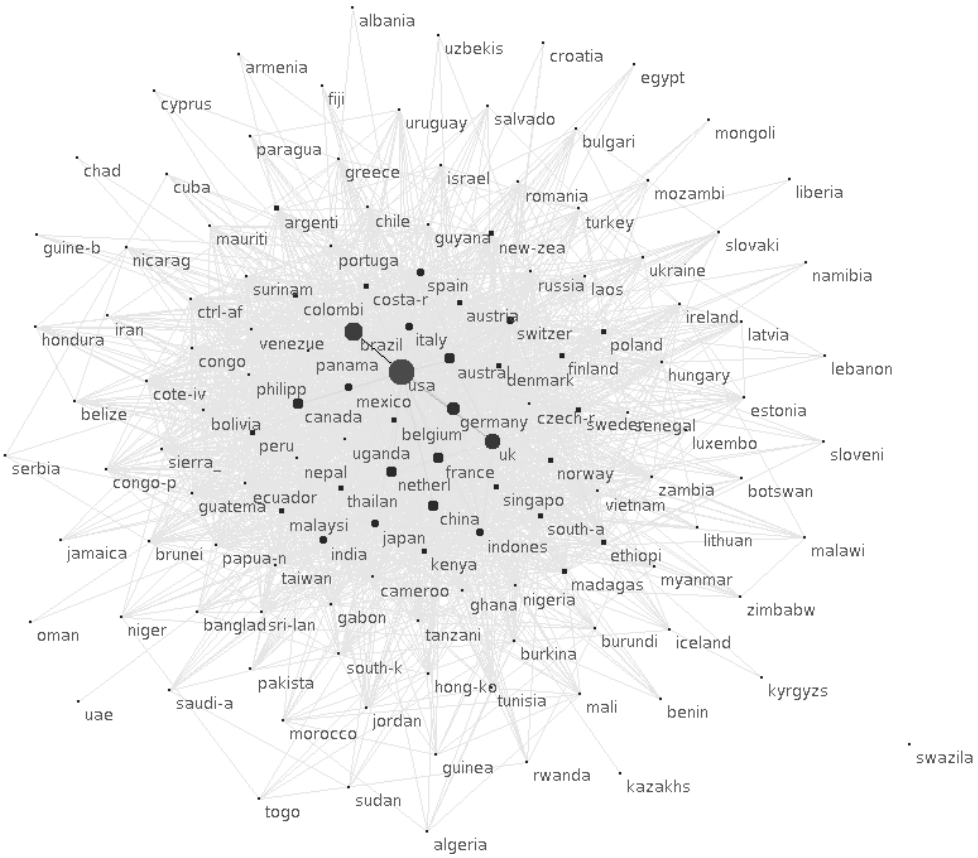

FIG. 5 – *Réseau formé par les pays publiant sur la déforestation pour les années 1975 à 2016.*

se révèle le pays ayant le plus souvent collaboré avec les États-Unis. Suivent ensuite le Royaume-Uni et l'Allemagne. Il est possible que la position très élevée de ces derniers dans le classement des pays publiant le plus sur la déforestation soit en partie due au fait de leur très forte collaboration avec des chercheurs des États-Unis. En effet Malhado et al. (2014) a démontré que la proportion d'articles par des auteurs de la région amazonienne sur l'Amazonie, particulièrement Brésiliens, a augmenté avec le temps mais que par ailleurs la proportion d'articles sur l'Amazonie n'impliquant pas d'auteurs de la région a également augmentée. Il peut s'avérer que même en augmentant considérablement leur participation à la production scientifique sur la déforestation, les Brésiliens n'arrivent pas forcément à en prendre le leadership.

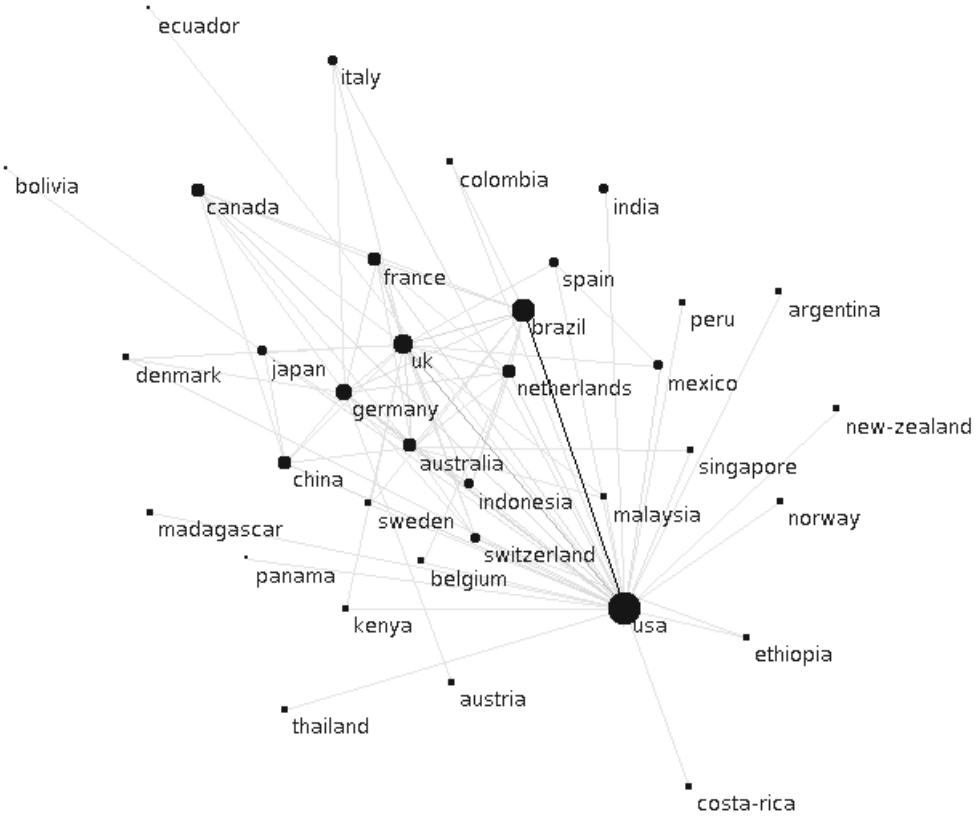

FIG. 6 – *Réseau de collaboration autour des États-Unis. Pour la période allant de 1975 à 2016.*

3.3 Nombre de publications par rapport au produit intérieur brut et la population en 2016

Le nombre de publications par habitant et par le produit intérieur brut exprimé en milliards de dollars américains a été calculé en utilisant les données fournies par la Banque Mondiale [2], pour l'année 2016.

	publi.	pop.	publi./pop	PIB'	publi./PIB'
brazil	236	207,65	1,14	1800	0,13
australia	112	24,12	4,64	1200	0,09
netherlands	72	17,01	4,23	771	0,09
uk	174	65,63	2,65	2620	0,07
canada	73	36,28	2,01	1530	0,05
germany	142	82,66	1,72	3470	0,04
france	86	66,89	1,29	2470	0,03
india	65	1324,17	0,05	2260	0, 03
usa	415	323,12	1,28	18600	0,02
china	91	1378,66	0,07	11200	0,01

TAB. 1 – *Nombre de publications pour chacun des 10 pays ayant le plus de publications par rapport à la population et au PIB (en milliards de dollars US) pour 2016. La première colonne "publi." représente le nombre de publications pour l'année 2016. La deuxième colonne "pop." représente la population en millions d'habitants. La troisième colonne "publi./pop" représente le ratio entre le nombre de publications et la population exprimée en million. La quatrième colonne "PIB'" représente le produit intérieur brut en milliards de dollars américains. La cinquième colonne "publi./PIB'" représente le nombre de publications divisé par le produit intérieur brut en milliards.*

Dans le tableau 1, nous voyons que l'Australie et les Pays-Bas ressortent comme étant les pays produisant le plus de publications par habitant. Dans le classement du nombre de publications en fonction du PIB, le Brésil arrive en tête suivi du duo Australie et Pays-Bas. Ces derniers voient peut-être leur leadership dans la recherche sur la déforestation limité par la taille de leur économie à l'échelle mondiale.

4 Sujets d'étude des publications

Les résumés des articles fournissent également des informations sur les pays, régions ou territoires auxquels les auteurs se sont les plus intéressés. Le tableau 2 fait ressortir ceux qui ont été le plus souvent mentionnés. L'Amazonie et le Brésil arrivent en tête, ce qui est un résultat attendu vu le nombre important de contributions Brésiliennes et le fait que la forêt Amazonienne est la plus importante au Brésil. Bien que les autres pays et régions soient moins souvent mentionnés, il est intéressant de constater que la quasi totalité des continents figure dans cette liste à l'exception de l'Océanie.

2. `https://data.worldbank.org/`

	publications
amazon	3863
brazil	2806
indonesia	956
africa	955
china	920
america	810
europe	785
mexico	635
costa rica	330
malaysia	248

TAB. 2 – *Les pays et régions les plus souvent mentionnés dans les résumés des publications pour la période de 1975 à 2016.*

4.1 Production par auteur et réseau de co-auteurs

Le graphe des co-auteurs dans la figure 7 fournit un aperçu sur les tendances de collaboration. Il montre les nombreux groupes formés par les auteurs qui collaborent sur le sujet.

Les auteurs ayant publié le plus figurent dans le tableau 3 avec un auteur Brésilien en tête, Fearnside, suivi d'un auteur d'Australie (Laurance) et d'un auteur Américain (Houghton) en deuxième et troisième positions respectivement. Deux auteurs sont donc issus des deux pays avec la plus importante production globale tandis que le troisième est issu d'un des pays avec la plus importante production par personne pour l'année 2016.

	publications
Fearnside, PM	97
Laurance, WF	73
Houghton, WF	63
Lambin, EF	58
Shimabukuro, YE	57
Koh, LP	56
Herold, M	49
Asner, GP	49
Achard, F	48
Peres, CA	44

TAB. 3 – *Publications des 10 auteurs ayant publié le plus.*

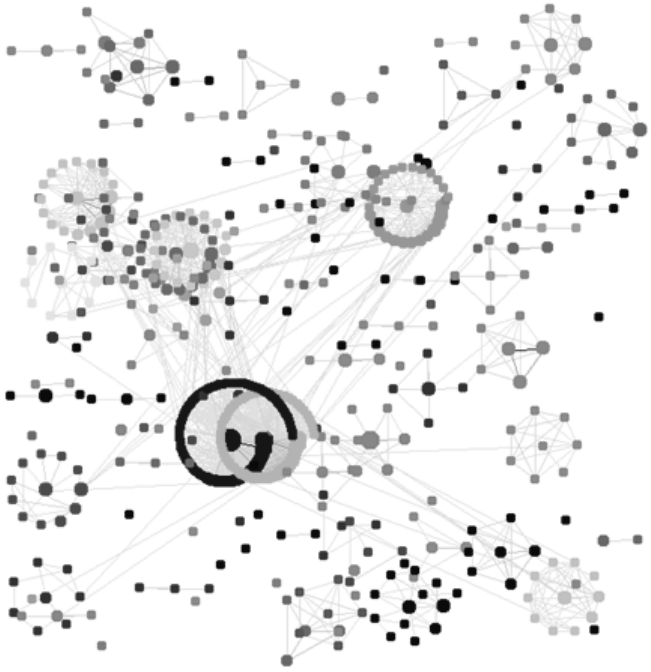

FIG. 7 – *Regroupement de co-auteurs.*

4.2 Les disciplines représentées

Le réseau des disciplines scientifiques les plus représentées est construit avec les catégories définies par le Web of Science (champ WC). Une publication pouvant se retrouver dans plusieurs catégories, il est possible de calculer les co-occurences des catégories et d'utiliser le résultat pour construire un réseau. Ainsi, il en ressort que la plupart des disciplines se regroupe autour de l'environnement et de l'écologie. Il se forme également un deuxième groupe ayant rapport à la médecine comprenant notamment la médecine tropicale et la parasitologie.

Le tableau 4 permet de voir les disciplines dans lesquelles le plus de publications ont été classées (champ WC du WoS). Ainsi, la majorité des publications est classée sous la catégorie des sciences environnementales. En regardant les autres disciplines il est possible de voir quelles sous-disciplines des sciences environnementales sont les plus représentées. L'écologie et les géosciences comptent le plus de publications. Il est possible de conclure que les publications sur la déforestation ont tendance à être interdisciplinaires, les sciences environnementales étant elles-mêmes interdisciplinaires et regroupant entre autre l'écologie et les géosciences.

	publications
Environmental Sciences	3890
Ecology	2812
Geosciences, Multidisciplinary	1571
Environmental Studies	1491
Biodiversity Conservation	1298
Forestry	1262
Meteorology & Atmospheric Sciences	1104
Geography, Physical	907
Remote Sensing	901
Multidisciplinary Sciences	807

TAB. 4 – *Les 10 catégories (champ WC du WoS) avec le plus de publications.*

5 Travaux reliés

Les analyses que nous avons présentées portent sur un corpus dans lequel se retrouvent les publications issues de diverses universités, laboratoires, unités de recherches et autres institutions. Les publications sont toutes liées au même thème de la déforestation.

Neptune (2014) a déjà utilisé des analyses bibliométriques de publications scientifiques pour analyser les activités de recherche au sein d'un unité scientifique spécifique, l'Institut de Recherche en Informatique de Toulouse. Ce travail portait sur toutes les publications présentes dans la base de donnée de l'unité de recherche, tous thèmes confondus. En utilisant les données sur l'organisation et le personnel du laboratoire, les analyses sur la production par équipe ainsi que sur les collaborations inter-équipe et la collaboration avec des auteurs extérieurs à l'unité ont pu être effectuées. L'auteur a montré comment l'analyse bibliométrique peut être utilisée pour certains aspects de l'évaluation d'une unité scientifique tels que la production scientifique, le rayonnement, l'implication dans la formation par la recherche et les perspectives scientifiques. Mothe et al. (2006) a démontré comment la plate-forme Tétralogie permet de combiner la fouille de données avec les fonctionnalités des systèmes d'information géographique pour découvrir la structure géographique d'un domaine. Les auteurs ont présenté une étude de cas en utilisant les actes de la conférence SIGIR (Special Interest Group on Information Retrieval) de l'ACM (Association for Computing Machinery). Les auteurs ont utilisé des cartes géographiques pour représenter visuellement la dimension géographique révélée par la fouille des données.

Les analyses présentées ici font suite à ces travaux. Nous utilisons la fouille de données et de méta-données de publications scientifiques pour analyser les activités de recherche sur le sujet de la déforestation. Nous nous sommes intéressés à la dimension géographique présente dans les données non seulement par rapport à la localisation des auteurs mais aussi par rapports aux zones et régions sur lesquelles portent leurs travaux de recherche.

Skupin (2014) a utilisé conjointement la bibliométrie et la visualisation des réseaux

pour faire ressortir la structure d'un domaine ainsi que les communautés basées sur les co-citations avec les publications de l'auteur David Mark. Cette approche a permit de réaliser une analyse visuelle de la dimension de l'influence de David Mark et sa persistance avec le temps dans le domaine des systèmes d'information géographiques.

Kang et al. (1990) a proposé une étude de faisabilité sur la méthode FODA (Feature-Oriented Domain Analysis) pour l'analyse d'un domaine. Cette méthode permet de créer un modèle du domaine en effectuant notamment une analyse de l'étendue du domaine. Les analyses présentées ici permettent d'élucider le domaine de la déforestation en vue de guider des travaux futurs sur les données liés à ce thème.

6 Perspectives

Les données collectées augmentées avec des données externes telles que celles de l'Organisation des Nations Unis pour l'alimentation et l'Agriculture et de la Banque Mondiale peuvent permettre de répondre à de nombreuses autres questions telles que celles liées à l'évolution des thèmes de recherche et à l'apport des institutions à chaque thème. Par exemple, le lien entre l'expérience de la déforestation dans les pays et la publication sur le sujet peut être examiné à partir de ces données. De plus, les collaborations entre institutions peuvent aussi être explorées avec des analyses de réseaux sociaux.

Ce travail pourrait être complété par une analyse plus poussée des résumés ou des articles complets notamment avec l'extraction des entités nommées qui faciliterait la mise en évidence des sujets spécifiques d'intérêts liés à la déforestation telles que des maladies, des plantes et des animaux spécifiques.

7 Conclusion

La fouille de texte des données et méta-données sur les publications scientifiques en rapport à la déforestation a permis d'avoir un aperçu de l'évolution de l'activité de recherche sur ce sujet au fil des années ainsi que la collaboration quasi-généralisée entre les pays et les nombreux réseaux de collaboration entre auteurs. Une augmentation quasi-régulière du nombre de publications est constatée d'une année à l'autre.

Le Brésil ressort comme un acteur important tant par la contribution de ses auteurs que par les mentions qui sont faites du pays dans les publications analysées.

Références

Bornmann, L. et R. Mutz (2015). Growth rates of modern science : A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology 66*(11), 2215–2222.

Diegues, A. C. S., P. Kageyama, et V. Viana (1992). *The social dynamics of deforestation in the Brazilian Amazon : an overview*, Volume 36. United Nations Research Institute for Social Development.

Dousset, B. (2009). Tetralogie : Software for monitoring science and technology.

Foley, J. A., R. DeFries, G. P. Asner, C. Barford, G. Bonan, S. R. Carpenter, F. S. Chapin, M. T. Coe, G. C. Daily, H. K. Gibbs, J. H. Helkowski, T. Holloway, E. A. Howard, C. J. Kucharik, C. Monfreda, J. A. Patz, I. C. Prentice, N. Ramankutty, et P. K. Snyder (2005). Global consequences of land us. *Science 309*, 570–574.

Kang, K. C., S. G. Cohen, J. A. Hess, W. E. Novak, et A. S. Peterson (1990). Feature-oriented domain analysis (foda) feasibility study. Technical report, Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst.

Malhado, A. C. M., R. S. D. de Azevedo, P. A. Todd, A. M. C. Santos, N. N. Fabré, V. S. Batista, L. J. G. Aguiar, et R. J. Ladle (2014). Geographic and temporal trends in amazonian knowledge production. *Computers, Environment and Urban Systems 46*, 6–13.

Mothe, J., C. Chrisment, T. Dkaki, B. Dousset, et S. Karouach (2006). Combining mining and visualization tools to discover the geographic structure of a domain. *Computers, Environment and Urban Systems 30*, 460–484.

Neptune, N. (2014). ánalyses bibliométriques des publications de l'irit.

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation 25*, 348–349.

Skupin, A. (2014). Making a mark: a computational and visual analysis of one researcher's intellectual domain. *International Journal of Geographical Information Science 28*(6), 1209–1232.

Summary

Deforestation is a widespread phenomenon that affects fairly large portions of land, especially in tropical regions. Remote sensing allows researchers to track and analyze the spatio-temporal evolution of this phenomenon.

Using text and metadata mining on scientific publications, on the theme of deforestation, we aim to identify the location of scientific production on deforestation and find out how researchers are connected to each other. Through network analysis, it is possible to highlight trends in terms of collaboration between authors. This network analysis reveals trends in the distribution of production among authors, whether it is concentrated at the level of particular authors in developed countries or whether it tends to be distributed in a balanced way between several developed and developing countries.

For this we rely on network analyses. Moreover, thanks to the analysis of the keywords we identify deforestation-affected sites that researchers are interested in, tropical forests and the Amazon, as well as related subjects related to the environment and the environment and health.