Les systèmes littéraires par le prisme de Wikipédia

Maxime Chabriel[*,1], Yasmine Houri[†,1], et Mathis Sansu[‡,1]

[1]ENSAE Paris, France

29 avril 2022

Résumé

Nous étudions l'apparition de genres littéraires francophones en lien avec les structures de réseau des systèmes littéraires sur Wikipédia. Nous construisons un graphe orienté dans lequel les noeuds correspondent aux auteurs francophones ayant une page Wikipédia à leur nom, et les arrêtes relient chaque auteur à ceux dont l'hyperlien est présent sur sa page Wikipédia. Nous construisons un deuxième graphe non orienté dans lequel les auteurs sont reliés par leur co-participation au genre. Nous discutons de plusieurs pondérations possibles pour chacun de ces réseaux. La comparaison de la structure de ces deux réseaux ne nous permet pas de conclure d'une similarité suffisamment significative entre les deux réseaux. Cependant, nous trouvons qu'une augmentation de la modularité du réseau d'auteurs une année donnée favorise l'émergence de nouveaux genres littéraires 45 ans plus tard. Malgré de nombreux biais, nous pensons que ces résultats peuvent être prometteurs pour une étude similaire avec des données plus fiables. Les données et scripts nécessaires à la réplication des analyses sont disponibles ici [1].

1 Introduction

L'étude de l'histoire littéraire éclaire traditionnellement l'évolution de la littérature à travers le prisme des courants littéraires et des relations entre littérature et histoire. Vaillant (2017) apporte une interprétation originale à cette discipline, en considérant la littérature comme un système culturel parmi d'autres. En définissant l'histoire littéraire comme l'histoire de la communication littéraire, il la fait entrer en résonance avec la sociologie et l'étude des réseaux. Il introduit alors la notion de "système littéraire", définie comme "le système socio-historique englobant l'ensemble des acteurs, des instances et des processus contribuant à l'histoire littéraire", qui émane de l'interaction de trois sphères : création, commentaire, et socio-économie. Comme cela a déjà été simulé théoriquement pour l'émergence de nouvelles disciplines scientifiques par Sun et al. (2013), nous nous intéressons à l'émergence de nouveaux genres littéraires, chaque genre étant une vision de l'écriture ou une manière d'écrire qui peut se définir par une communauté d'auteusr partageant cette vision ou empruntant des éléments du genre en question dans leur propre travail. L'émergence de nouveaux genres littéraires serait alors le résultat non seulement de la production littéraire elle-même, mais également des interactions entre auteurs, commentateurs et agents sociaux, enclavées dans les époques et les systèmes d'organisation sociale.

Afin de tester ces hypothèses, il conviendrait d'observer les interactions entre les auteurs au cours de l'histoire littéraire, et d'estimer dans quelle mesure les propriétés de ces interactions prédisent l'apparition de nouveaux genres littéraires. L'observation directe de ces interactions passées étant impossible, les données utilisées doivent être trouvées sur une source recensant les interactions littéraires, économiques et sociales intertemporelles entre les auteurs francophones, en lien avec les genres dans la littérature française. L'encyclopédie en ligne Wikipédia remplit ces critères. L'opérationnalisation de nos concepts, bien qu'imparfaite, repose donc sur le proxy constitué par les pages Wikipédia des auteurs francophones.

L'ambition de notre travail est d'éclairer l'apparition de genres littéraires par les structures de réseau des systèmes littéraires selon une dimension historique, tels que mis en avant par l'architecture d'hyperliens de Wikipédia. Cette démarche englobe deux questions sous-jacentes :

— Est-ce que le réseau d'auteurs construit à partir de la présence des hyperliens sur leurs pages respectives s'identifie au moins partiellement au réseau d'auteurs construit à partir de la co-participation à un genre ?

— Peut-on rendre compte de l'émergence d'un genre à partir de la structure du réseau des auteurs sur Wikipédia ?

[*]maxime.chabriel@ensae.fr

[†]yasmine.houri@ensae.fr

[‡]mathis.sansu@ensae.fr

1. https://github.com/ChabiMax/author_network

2 Données

Les données utilisées sont obtenues par aspiration automatique du site Wikipédia. Nous fondons notre collecte de données sur la liste des auteurs francophones par ordre chronologique [2] et la liste des genres littéraires en langue française [3, 4].

Nous extrayons quand elles sont disponibles les informations suivantes sur les auteurs : nom et prénom, date de naissance, date de mort, hyperlien de la page Wikipédia, longueur du code source de la page. Dans un second temps, nous ramenons dans notre base de données les informations sur les liens avec d'autres auteurs : identifiants des auteurs mentionnés sur la page, nombre de fois où chaque auteur est mentionné, identifiant des genres mentionnés, identifiant des auteurs contemporains, différence d'âge avec les auteurs contemporains, nombre d'années durant lesquelles les deux auteurs étaient en vie à la même époque. Enfin, nous définissons les dates de début et de fin d'un genre respectivement par le minimum des dates de naissance et maximum des dates de mort des auteurs contribuant au genre [5]. Nous construisons deux tables d'adjacence - et donc deux réseaux - à partir de ces données, dont nous détaillons les procédures dans la section suivante.

3 Méthodes

3.1 Construction des réseaux

Pour répondre à nos problématiques, nous construisons deux réseaux différents. Le premier (ci-après G_A), modélise le réseau des auteurs liés par la mention de l'hyperlien de l'un sur la page Wikipédia de l'autre (réseau dirigé de 1584 dont 1250 non isolés, 13021 arêtes, 6421 cliques). Son coefficient de transitivité est proche de $0,3$ (sur l'ensemble des triplets de noeuds connectés, 33% sont des triades fermées), le coefficient de réciprocité globale est proche de 0.5 (en moyenne, un auteur est réciproquement mentionné sur la moitié des pages des auteurs qu'il mentionne lui-même). Le deuxième réseau (ci-après G_G) que nous construisons relie les auteurs dont la page Wikipédia mentionne au moins un genre littéraire en commun (réseau non dirigé, de 1584 noeuds dont 731 connectés). Comme certains genres littéraires sont extrêmement généraux (roman, poésie, *etc.*) le coefficient de transitivité du réseau est élévé, à $0,93$.

3.2 Contrôle du nombre de connexions par la taille des articles

La relation de l'auteur i à l'auteur j par la présence d'un hyperlien vers j dans la page de i est pondérée par le nombre de fois où l'hyperlien de j est trouvée dans la page de i : ce poids est noté $w_{i,j}$. Néanmoins, le nombre de connexions d'un auteur est susceptible d'être surestimé selon le zèle des contributions sur Wikipédia : plus un article est long, et plus la probabilité qu'un lien vers un autre auteur apparaisse est forte [6]. Pour déterminer l'influence de la taille X_i d'un article i sur son nombre de connexions Y_i, nous effectuons la régression linéaire suivante :

$$Y_i = \beta log(X_i) + \epsilon_i \tag{1}$$

Notons bien que la constante usuellement introduite dans les modèles de régression linéaire est ici fixée à 0, car la relation $f(.)$ entre la taille d'un article x et le nombre $f(x)$ d'hypeliens qu'il inclut en son sein ne fait aucun sens si $f(0) \neq 0$: aucun auteur ne peut être mentionné dans un article de taille nulle. Les résidus ϵ_i de (1) correspondent alors au nombre "réel" de connexions que l'auteur possède avec d'autres auteurs - *via* les hyperliens présents sur sa page -, c'est à dire le nombre de connexions retranché de l'effet artificiel de la taille de l'article. Pour éviter de faire disparaître certains liens, nous assignons la valeur 0.1 aux résidus négatifs. Les nouveaux poids $w'_{i,j}$ sont construits de façon à ce que leur somme soit égale aux ϵ_i pour chaque individu. On a donc :

$$w'_{i,j} = w_{i,j} \times \frac{max(\epsilon_i,\, 0.1)}{\sum_j w_{i,j}} \tag{2}$$

2. https://www.wikipedia.org/wiki/Liste_d%27%C3%A9crivains_de_langue_fran%C3%A7aise_par_ordre_chronologique
3. https://fr.wikipedia.org/wiki/Genre_litt%C3%A9raire
4. Collecte des données achevée le 26 avril 2022.
5. Dans le cas où des auteurs ont un genre mentionné sur leurs pages et sont encore en vie, nous assignons l'année 2022 comme fin du genre.
6. Nous faisons cette hypothèse, car il apparaît probable que les contributions sur le Wikipédia des auteurs francophones ne reflètent pas exactement l'importance réelle dans le système littéraire d'un auteur, en même temps que ces contributions sont dépendantes des connaissances historiques sur le champ littéraire.

Une pondération similaire est appliquée au réseau de co-participation au genre, où les $w_{i,j}$ mesurent le nombre de genres littéraires que les auteurs i et j possèdent en commun. Une différence tient cependant au fait que le réseau est non dirigé (les liens reposent sur la co-participation) et donc la pondération doit prendre en compte les attributs de chaque noeud du lien, à savoir les tailles des deux articles des auteurs. On calcule le coefficient de pondération pour chacun des noeuds comme s'il produisait des liens dirigés. Chaque lien a donc deux coefficients de pondération qui lui sont rattachés selon la taille de l'article de chaque auteur du lien. Nous assignons comme poids au lien la moyenne de ces deux coefficients de pondération.

3.3 Pondération des genres littéraires

La définition des genres littéraires à partir de leur page Wikipédia se confronte à de nombreux biais, liés au système d'organisation de l'encyclopédie en ligne. En particulier, certains genres sont imbriqués les uns dans les autres, et notre collecte de données par aspiration automatique ne nous permet pas d'en rendre compte : par exemple, le roman et le roman historique sont comparés sur un même niveau. Cependant on ne peut pas non plus les subsumer les uns dans les autres, car cela créerait de nouveaux biais : quels genres subsumer, à quel niveau ? De plus, tous les genres ne sont pas exactement imbriqués les uns dans les autres, et les recoupements se font dans un espace continu. Pour surmonter cet obstacle, nous faisons la remarque suivante : si un genre littéraire peut se définir par une communauté d'auteurs (*cf.* Introduction), la proximité entre deux genres peut alors se mesurer par le nombre d'auteurs communs entre deux genres. Pour un genre i, on définit un poids w_i tel que :

$$w_i = \prod_j \frac{1}{2} \times \left(1 + \frac{\#\{auteurs\ dans\ i\ ou\ j\}}{\#\{auteurs\ dans\ i\} \ +\ \#\{auteurs\ dans\ j\}}\right) \tag{3}$$

Dans le réseau G_G, il faut pouvoir exprimer cette pondération des genres sur le poids des liens entre deux auteurs. Le poids du lien entre les noeuds i et j est alors redéfini comme le produit du coefficient de pondération des genres littéraites et du coefficient de podération selon la taille des articles (*cf.* sous-section précédente).

3.4 Procédure d'analyse pour répondre aux problématiques posées

Pour comparer les réseaux G_A et G_G, nous établissons des métriques pour observer si les relations d'un auteur évoluent d'un réseau à l'autre, et le cas échéant, de quelle manière. Nous explorons entre autre son nombre de connexions. Pour déterminer une relation temporelle entre les réseaux, où une certaine structure du réseau G_A déterminerait l'apparition d'un nouveau genre, nous faisons appel à un découpage de G_A en des sous-réseaux $G_{A,t}$, où ne sont conservés que les noeuds correspondants à des auteurs alors vivants et ayant au moins 20 ans l'année t. Nous extrayons ensuite plusieurs métriques pour chacun des $G_{A,t}$, et les utilisons dans une régression aux moindres carrés ordinaires (MCO) pour prédire l'apparition d'un nouveau genre.

4 Résultats

4.1 Comparaison des structures des réseaux

Dans un premier temps, nous comparons les degrés des noeuds dans G_A et G_G, afin d'évaluer si le nombre d'auteurs avec lesquels un auteur i est lié par hyperlien est corrélé au nombre d'auteurs qui co-participent à au moins un même genre que i. Pour ce faire, nous calculons le degré de chaque noeud dans chacun des réseaux, les classons par odre décroissant d'importance, et observons l'intersection des n auteurs les plus importants dans chaque réseau avec les n auteurs les plus importants dans l'autre. La Figure 1 est une *heatmap* représentant la part des x noeuds les plus importants de G_G dans les y noeuds les plus importants de G_A en fonction du nombre x de noeuds retenus. Ainsi, on observe que le noeud le plus important dans G_A est présent dans tous les groupes de noeuds les plus importants dans G_G de taille supérieure ou égale à 6. On y lit également que 50% des 12 noeuds les plus importants dans G_A sont présents dans les groupes de noeuds les plus importants dans G_G de taille supérieure ou égale à 36. Il n'apparaît donc pas qu'il y ait de corrélation significative entre avoir une position de centralité dans G_A et en avoir une dans G_G.

Dans un second temps, nous observons graphiquement la corrélation entre le nombre de connections d'un auteur dans le réseau G_A et son nombre de connections dans le réseau G_G. La Figure 2 représente le degré dans G_A en fonction du degré dans G_G sous forme de nuages de points, augmentés d'une régression linéaire. La pente de la régression rend compte d'une corrélation positive entre le degré de chaque noeud dans G_A et dans G_G, mais cette corrélation ne nous permet pas de conclure d'une franche similarité entre ces deux réseaux.

4.2 Émergence d'un genre littéraire à l'aune des structures de réseau d'auteurs

Le Wikipédia francophone recense au total 202 genres littéraires, on en retrouve 134 dans les pages biographiques des auteurs francophones. Pour se ramener à une variable d'étude continue, nous utilisons un lissage polynomial de la fonction $f(t) = y$ où y désigne le nombre de nouveaux genres apparus lors de l'année t. L'année d'apparition d'un genre est déterminée par la procédure suivante : de toutes les dates de naissance des auteurs faisant partie du genre, on fait la moyenne des deuxième et troisième plus basses (voir Figure 3). Comme la date de naissance d'un auteur ne correspond pas à son *année d'entrée* dans un genre littéraire, on rajoute ensuite 30 ans à cette variable. Cette procédure permet d'introduire davantage de robustesse à notre variable dépendante, dans la mesure où les valeurs extrêmes d'une distribution sont usuellement très sensibles aux variations.

A la suite des analyses précédentes, nous rendons compte des corrélations entre structures de réseaux et émergence de genres analytiquement. Visuellement, il est difficile de se prononcer sur l'existence de corrélation entre l'apparition de nouveaux genres et l'*average clustering*[7] des sous-réseaux $G_{A,t}$. Cependant, la modularité nous apparaît prometteuse (Figure 4). Nous adoptons le modèle de régression linéaire MCO avec comme variable dépendante le nombre de nouveaux genres, expliqué par la modularité (chacune des variables appliquant le système de pondération telle que décrit dans la section 3.2.) retardée d'un nombre d'année h. Cette régression est contrôlée par la somme du logarithme de la taille (en caractères) du code html des pages de chaque auteur vivant de plus de 20 ans en t[8], ainsi que l'année t (en effet, la probabilité qu'un nouveau genre littéraire apparaisse dans un corpus de page Wikipedia dépend mécaniquement en partie de la taille de ce corpus). Les résultats de cette régression sont reportés dans les Figures 6 (pour un retard de $h = 45$ ans) et 7. On trouve que la variable modularité a un coefficient le plus élevé dans la régression lorsque spécifié avec un retard de $h = 45$. Autrement dit, une modularité élévée à une année t a tendance à faire émerger de nouveaux genres littéraires en $t + 45$.

5 Discusssion

De la même manière que pour les réseaux et les disciplines scientifiques (Sun et al. (2013)), nous trouvons une corrélation positive entre le niveau de modularité d'un réseau et l'apparition d'un genre littéraire. Si l'on suit le même raisonnement que les auteurs de l'article cité plus haut, l'isolement et la cohésion simultanée d'un groupe d'auteurs du reste de la communauté littéraire font qu'à l'échelle d'une vie (45 ans), de nouvelles façons d'écrire, ou de percevoir l'écriture et la littérature, se dessinent, faisant apparaître un nouveau genre littéraire.

Cependant, notre opérationnalisation des réseaux présente des limites liées à la nature de notre source de données. Comme précisé en introduction, le jeu de données parfait à utiliser pour notre analyse aurait résulté de l'observation directe des systèmes littéraires et de leur évolution au cours du temps, mais cette observation est impossible. Du fait de notre choix d'utiliser Wikipédia comme source de données, nous n'observons pas réellement les systèmes littéraires eux-mêmes, mais leur reconstitution par les contributeurs de cette encyclopédie et par l'aggrégation des pages retenues. De ce fait, nos données sont soumises à de nombreux biais liés aux choix des contribtueurs, à la structure de Wikipédia, et aux limites de notre connaissance des faits historiques passés.

Par ailleurs, bien que nous ayons décidé de prendre en compte dans la pondération l'effet de la taille de l'article sur la taille du réseau en supposant que plus un article est long, plus le nombre d'auteurs qui y sont cités est suceptible d'être grand, il peut en réalité y avoir une causalité inverse entre ces deux variables : la taille réelle (inobservable) d'un réseau a probablement un effet direct sur la longueur des pages Wikipédia qui lui sont associées.

Finalement, si nous sommes capables de trouver une corrélation entre la modularité du réseau des auteurs littéraires français et l'émergence d'un genre 45 ans plus tard, nous ne sommes pas en mesure, avec nos données telles qu'elles nous se sont présentées, de retracer individuellement un genre à un groupe d'auteurs particulier. Encore une fois, notre étude pourrait beaucoup gagner d'un réseau d'auteurs tracé à la main, par des experts en littérature française.

Références

Sun, X., Kaur, J., Milojević, S., Flammini, A., and Menczer, F. (2013). Social Dynamics of Science. *Scientific Reports*, 3(1) :1069.

Vaillant, A. (2017). *L'histoire littéraire*. Collection U. Armand Colin, Paris, 2e éd. revue et augmentée edition.

7. Moyenne des coefficients locaux de clusterisation.
8. On se réfère à la Figure 5 pour visualiser pourquoi il peut être judicieux d'ajouter ce contrôle.

Annexes

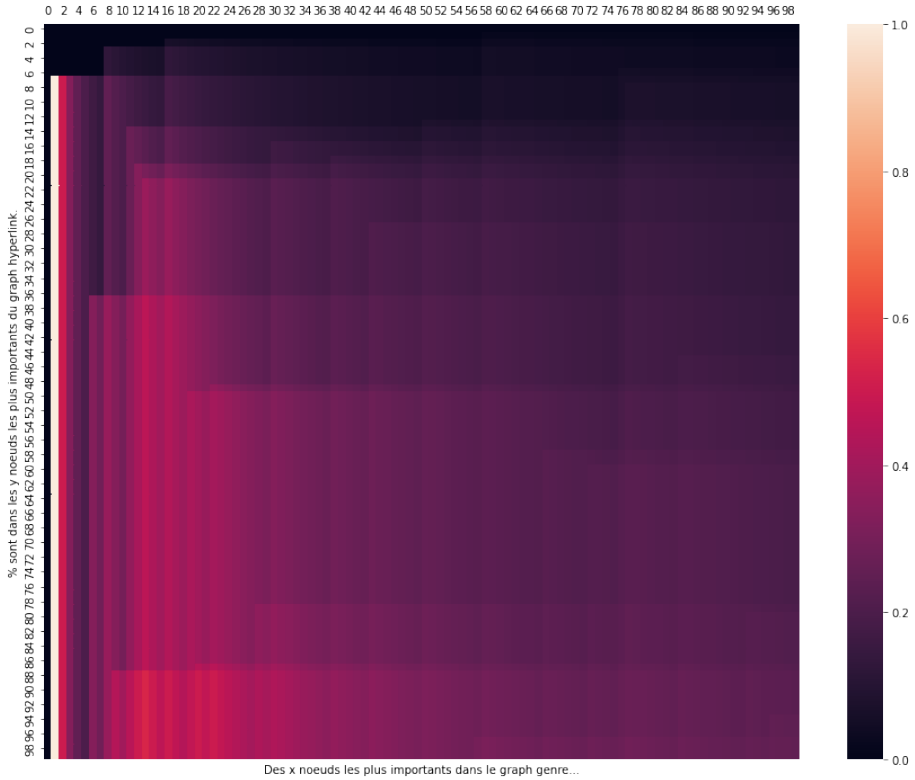

FIGURE 1 – Part des noeuds les plus importants de G_A dans les noeuds les plus importants de G_G.

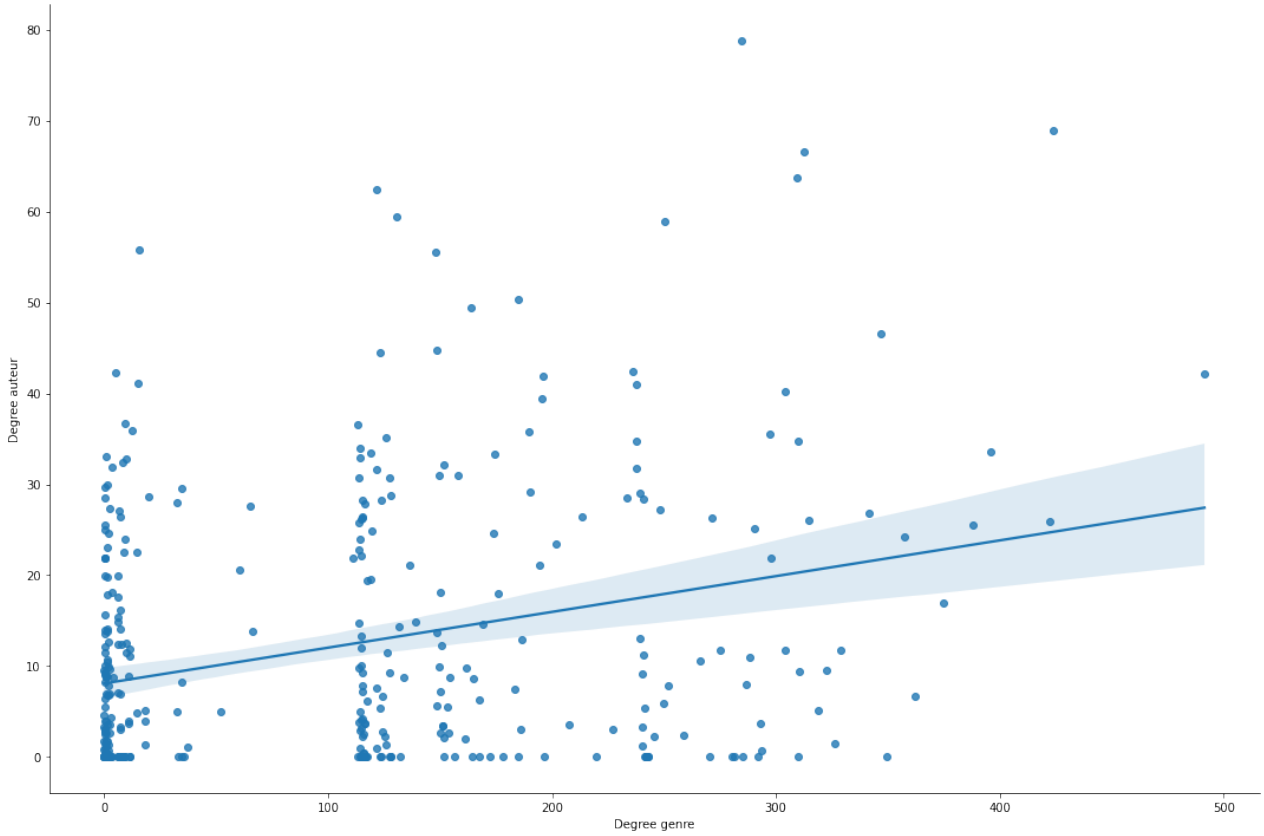

FIGURE 2 – Pour chaque auteur, nombre de liens dans G_A (ordonnée) et nombre de liens dans G_B (abscisse).

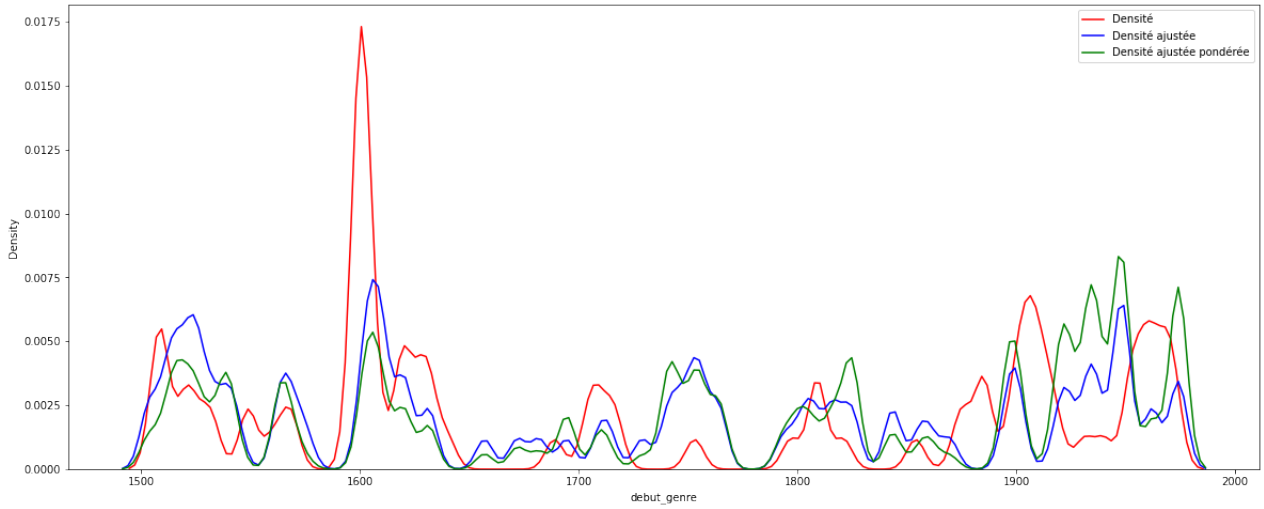

FIGURE 3 – Densité d'émergence des genres avec ou sans la pondération, avec ou sans le lissage sur la date d'émergence.

FIGURE 4 – Densité d'émergence des genres comparée à l'évolution de la modularité (définie avec ou sans pondération).

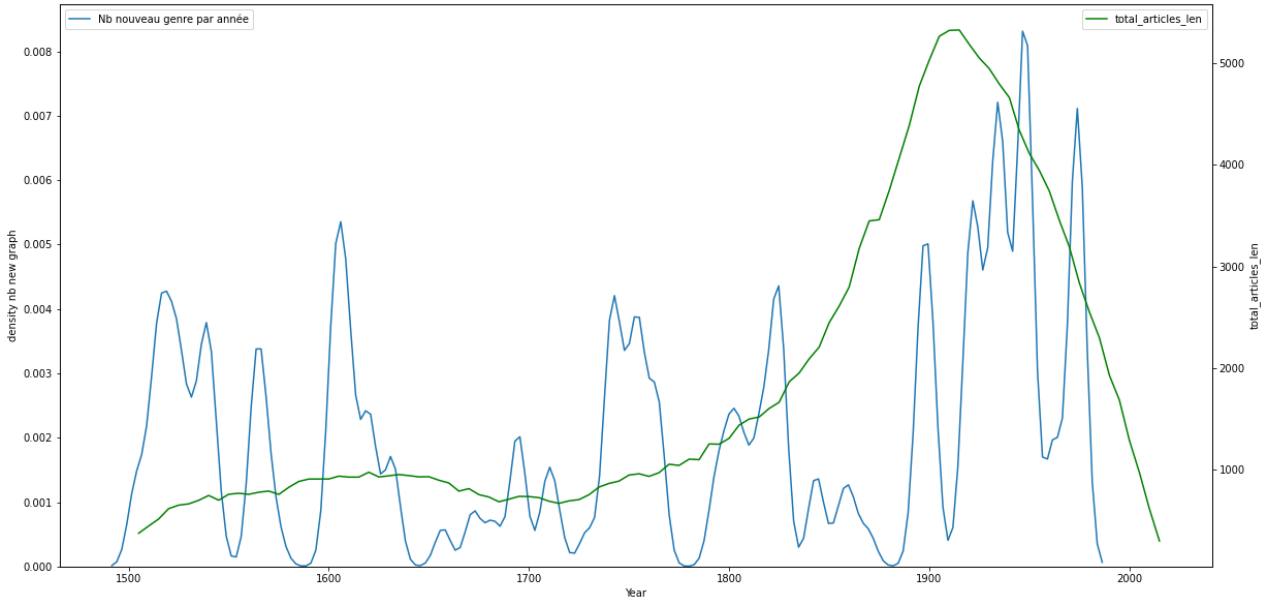

FIGURE 5 – Densité d'émergence des genres comparée à la somme du logarithme de la taille du html (en charactères) de l'ensemble des article Wikipédia des auteurs de plus de 20 ans vivants une année donnée.

OLS Regression Results

Dep. Variable:	annee_mean_interpolated	R-squared:	0.281
Model:	OLS	Adj. R-squared:	0.255
Method:	Least Squares	F-statistic:	10.94
Date:	Fri, 29 Apr 2022	Prob (F-statistic):	3.88e-06
Time:	21:16:43	Log-Likelihood:	-147.76
No. Observations:	88	AIC:	303.5
Df Residuals:	84	BIC:	313.4
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.1219	3.184	-0.980	0.330	-9.454	3.210
modularity_weight2	4.1628	1.575	2.644	0.010	1.031	7.294
total_articles_len	0.0002	0.000	1.438	0.154	-9.26e-05	0.001
annee	0.0018	0.002	0.958	0.341	-0.002	0.006

Omnibus:	4.805	Durbin-Watson:	0.451
Prob(Omnibus):	0.090	Jarque-Bera (JB):	3.473
Skew:	0.343	Prob(JB):	0.176
Kurtosis:	2.311	Cond. No.	6.60e+04

FIGURE 6 – Sortie de régression de l'année d'apparition d'un genre sur la modularité pondérée, contrôlée par la taille de l'article et l'année d'estimation. La modularité est laggée de 45 ans.

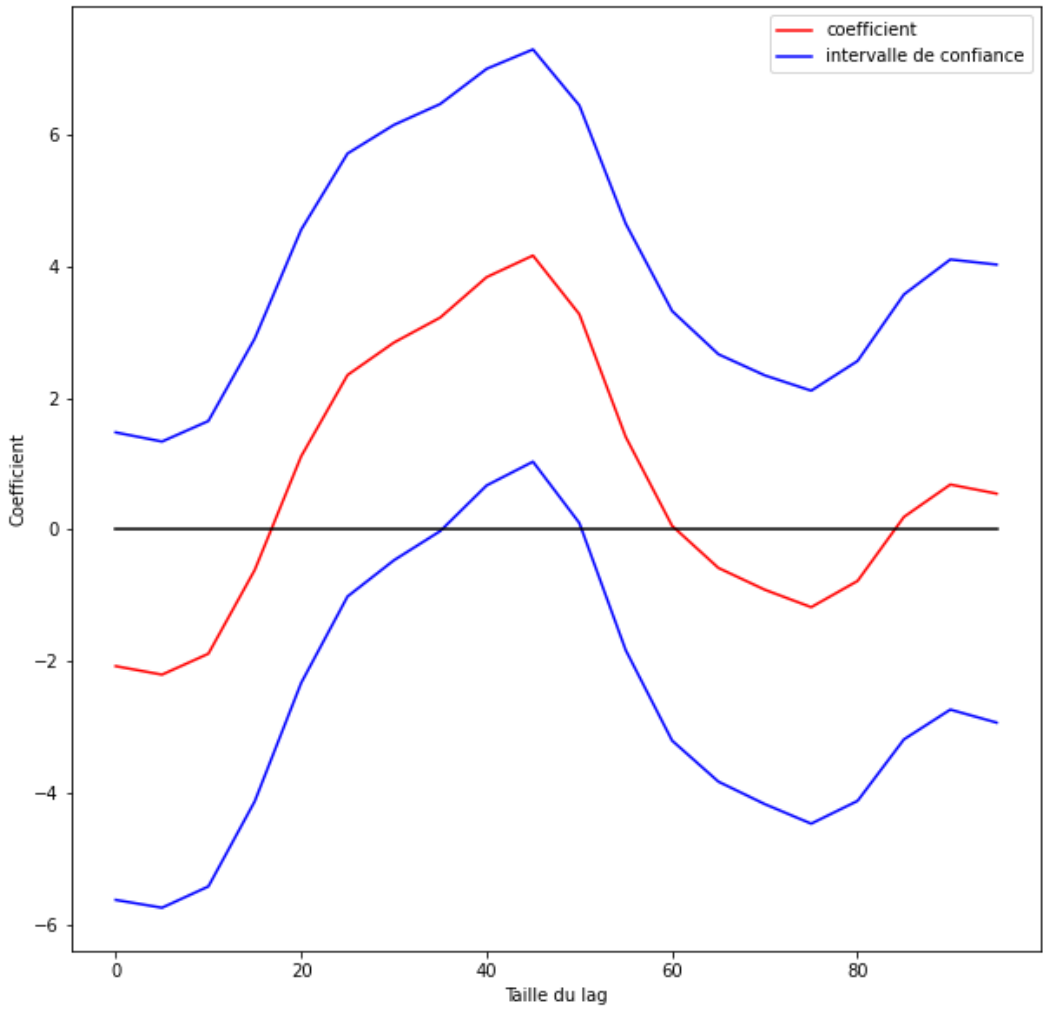

FIGURE 7 – Coefficients associés à la modularité en tant que variable explicative dans la régression sur nombre de nouveaux genres, selon le retard appliqué à la modularité dans la spécification.