Auteur : Romain JAFFUEL

Analyse de l'Engagement des Apprenants dans les MOOCs : Impact du Genre et de l'IDH

Cursus : Data & Humanité Digitales

Sciences Po Saint-Germain-en-Laye
CY Tech

Superviseur : Docteur **Matthieu Cisel**

Janvier 2025

Résumé

Cette étude s'est focalisée sur l'analyse des comportements d'engagement des apprenants dans des MOOCs, en examinant l'influence des facteurs sociodémographiques, tels que le genre et l'indice de développement humain (IDH). L'objectif principal était d'identifier les déterminants majeurs de l'engagement, mesuré par le nombre de vidéos visionnées et les quiz réalisés. Les données utilisées provenaient de trois itérations d'un MOOC, collectées sous forme de logs d'activité et de réponses à des enquêtes. Une approche quantitative a été adoptée, incluant des tests de Student et des régressions logistiques pour évaluer les effets principaux et les interactions entre variables. Les résultats ont montré que le IDH a un effet significatif sur le nombre de vidéos visionnées, avec une augmentation progressive associée à des contextes plus favorables. En revanche, le genre n'a pas présenté d'effet significatif. Ces résultats suggèrent que les disparités d'engagement sont davantage influencées par les contextes socio-économiques que par les différences liées au genre. Ces observations soulignent la nécessité de concevoir des interventions éducatives adaptées aux besoins spécifiques des apprenants provenant de contextes moins favorisés.

Table des matières

Liste des Figures

Liste des Tables

Introduction

Depuis leur apparition, les cours en ligne ouverts et massifs (MOOCs) ont révolutionné l'accès à l'éducation en permettant à des millions d'apprenants à travers le monde de se former à moindre coût et sans barrières géographiques. Toutefois, ces plateformes doivent faire face à un problème récurrent : des taux d'abandon élevés, souvent attribués à un faible engagement des apprenants ou à des obstacles liés à la conception des cours. Selon une étude récente (Wintermute et al, 2021) sur la plateforme de MOOC « France Université Numérique », la proportion moyenne des inscriptions aboutissant à une certification est inférieure à 10 %. Ces faibles résultats soulèvent des questions cruciales sur la manière d'améliorer l'expérience et les résultats des apprenants dans les MOOCs.

Malgré de nombreuses études sur les déterminants de l'abandon dans les MOOCs, il subsiste un manque de connaissances sur les interactions entre facteurs d'engagement des apprenants et conception des cours. La littérature actuelle ne permet pas de pleinement comprendre comment certains paramètres – tels que le nombre de cours suivis simultanément ou les différences d'engagement entre groupes d'apprenants – influencent directement les résultats d'apprentissage. Cette lacune en termes de travaux empiriques limite les capacités des concepteurs à proposer des interventions personnalisées.

Quels sont les effets des différentes catégories d'apprenants sur les taux de certification dans les MOOCs ? Comment le genre, l'indice de développement humain (IDH) et les pratiques des profils comportementaux modulent-ils les résultats ? Ces caractéristiques influencent-elles la réussite de manière significative et reproductible d'une itération à l'autre selon les itérations du cours ?

Pour explorer ces questions, nous avons analysé un jeu de données comprenant les logs et les réponses à des questionnaires collectés sur trois itérations d'un MOOC (15182 apprenants). Ces données anonymes ont été consolidées pour permettre une étude comparative entre les différents groupes d'apprenants définis selon leur comportement. Notre approche est quantitative au regard de la grande quantité de données que l'on traite et de la méthode de collecte automatique. Des techniques statistiques telles que le test de Student et les tests de corrélation ont été mobilisés pour évaluer les effets des variables d'intérêt.

Nous postulons que les apprenants les plus engagés obtiennent des résultats nettement supérieurs aux autres catégories, et que les facteurs sociodémographiques tels que le genre ou l'IDH influencent significativement ces résultats. Cette hypothèse repose sur des études précédentes montrant que l'engagement est un prédicteur majeur de la réussite dans les MOOCs. Les analyses suivantes visent à valider ou à nuancer ces suppositions, en fournissant des pistes pour améliorer les designs pédagogiques adaptés aux besoins des apprenants. Pour examiner ces hypothèses et répondre aux questions soulevées, une méthodologie rigoureuse a été mise en place, reposant sur l'analyse de données quantitatives issues de plusieurs itérations d'un MOOC, comme détaillé dans la section suivante.

Méthodes et données

Les données utilisées dans cette étude proviennent de Unow, une plateforme française de MOOC. Les données sont des rapports d'activité, des carnets de notes, et des réponses aux enquêtes des participants à deux cours en ligne, désignés comme MOOC1 et MOOC2. Ces informations ont été téléchargées directement depuis les plateformes de gestion des cours, garantissant ainsi leur authenticité et leur exhaustivité. Pour préserver la confidentialité des utilisateurs, toutes les données étaient entièrement anonymisées avant leur analyse, conformément aux politiques de confidentialité

et aux conditions d'utilisation des plateformes. Les données relatives aux pays de résidence des participants ont été obtenues à partir des réponses aux enquêtes de mars à juillet 2014. Les indices de développement humain associés aux pays des participants ont été extraits des données officielles des Nations Unies.

Une vérification des jeux de données a été effectuée pour identifier les valeurs aberrantes et anomalies potentielles. Les vidéos ne faisant pas partie du contenu pédagogique des cours, telles que les tutoriels ou les introductions hebdomadaires, ont été exclues des analyses, car elles ne reflètent pas l'activité d'apprentissage. Les étudiants ayant la possibilité d'obtenir des crédits académiques pour leur participation ont également été retirés des analyses, puisqu'ils ne correspondaient pas au profil visé de participants en auto-apprentissage. Ces étapes ont permis de garantir la validité des analyses en limitant les biais liés à des comportements atypiques ou à des données non pertinentes.

Les données ont été consolidées à partir de plusieurs fichiers contenant les logs et les résultats des enquêtes des deux MOOCs. Ces fichiers ont été fusionnés afin de constituer une base de données cohérente et homogène, regroupant toutes les itérations disponibles. On a aussi créé des variables dérivées, telles que le nombre total de vidéos visionnées ou de quiz soumis, afin de faciliter les analyses ultérieures. La suppression des variables fortement corrélées a été réalisée après une analyse exploratoire, incluant la création de corrélogrammes pour identifier les relations entre les variables.

Une première étape a consisté à évaluer la qualité des données. Cette analyse a révélé un faible taux de données manquantes pour les variables clés, mais des disparités notables existent entre certaines catégories.

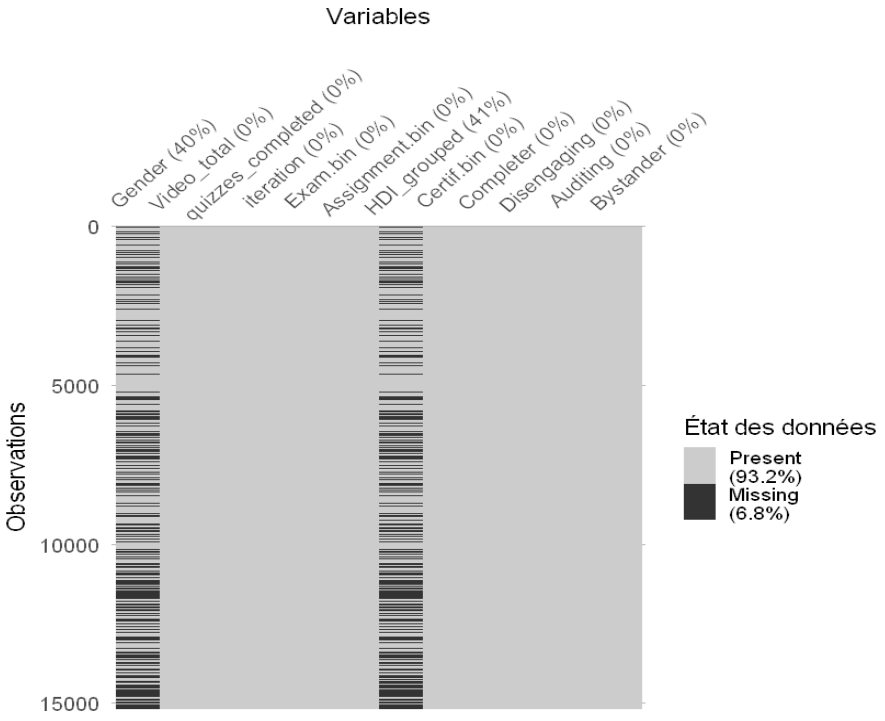

Figure 1 : Carte de chaleur des données manquantes.

Cette visualisation met en évidence les données manquantes de différentes variables. Par exemple. Les valeurs manquantes sont pour les données de genre à 40% et d'indice de développement humain à 41%.

L'analyse des données a été réalisée avec le logiciel R 4.4.1. Les principaux outils mobilisés incluent des packages dédiés au nettoyage et à la manipulation des données, tels que dplyr et tidyr, ainsi que des outils graphiques comme ggplot2 pour les visualisations. L'analyse des données manquantes a été effectuée à l'aide de naniar. Ces choix méthodologiques et techniques s'inscrivent dans des pratiques largement reconnues et utilisées dans des études similaires, assurant ainsi la robustesse et la reproductibilité des résultats.

Les participants ont été catégorisés selon leur niveau d'engagement, reflétant des degrés de participation distincts au cours. Ceux ayant obtenu un certificat ont été qualifiés de *Completer*, marquant une implication complète. Les étudiants ayant soumis au moins un quiz ou un devoir sans pour autant achever le cours ont été désignés comme *Disengaging*. Parmi ceux n'ayant soumis ni quiz ni devoir, une distinction a été faite entre les *Auditing*, qui ont visionné au moins 10 % des vidéos disponibles, et les *Bystander*, dont l'engagement est resté en dessous de ce seuil (Anderson et al., 2014). Bien que le terme *Disengaging* puisse être débattu, car la soumission d'un quiz n'illustre pas nécessairement un engagement profond, cette étape représente une implication plus significative que le simple visionnage de vidéos, comme l'ont démontré cet article et d'autres travaux (Ho et al., 2014). La distribution des différentes catégories est représentée dans la table suivante :

TABLE 1 – Distribution des catégories d'étudiants par itération du MOOC

Iteration	Completer	Disengaging	Auditing	Bystander
1	31.1	27.6	0.7	39.4
2	27.0	25.4	1.5	44.8
3	25.7	21.8	0.9	50.0

On observe une diminution progressive de la proportion des *Completers* (de 31,1% à 25,7%) et des *Disengaging* (de 27,6 % à 21,8 %), tandis que la proportion des *Bystanders* augmente significativement, passant de 39,4 % à 50.0 %. Ces tendances reflètent une réduction de l'engagement global au fil des itérations, accompagnée d'une hausse du désengagement et d'une plus forte proportion d'apprenants passifs (*Bystanders*).

Ces analyses exploratoires ont permis de préparer le jeu de données en mettant en évidence des dynamiques d'engagement différenciées parmi les apprenants des MOOCs. À partir de cette base, les résultats suivants s'intéressent aux liens entre les catégories d'apprenants, leurs comportements, et les facteurs sociodémographiques afin de mieux comprendre les mécanismes sous-jacents à l'engagement et à la réussite dans ces cours en ligne.

Résultats

Cette section présente les principaux résultats de l'analyse, en s'appuyant sur des visualisations pour illustrer les tendances et relations observées. Nous discutons d'abord de la qualité et des caractéristiques des données, avant d'aborder les résultats issus des modèles statistiques.

1.1 Analyse des comportements d'engagement en fonction des facteurs sociodémographiques et des interactions

1.1.1 Comparaison du nombre de vidéos vues selon le genre

Nous avons dans un premier temps comparé le nombre moyen de vidéos visionnées entre les genres à l'aide d'un test de Student, en supposant une distribution normale de la variable. Les résultats indiquent une différence significative ($t(5879, 2) = -3.76$, $p < 0.001$), les femmes ayant visionné en moyenne 14.1 vidéos, contre 13.1 pour les hommes. Étant donné que la distribution n'est pas normale, un test non-paramétrique de Wilcoxon a également été effectué. Ce dernier a confirmé une différence significative ($W = 8712151$, $p < 0.001$). Ces analyses suggèrent que les femmes ont, en moyenne, un engagement légèrement supérieur en termes de visionnage de vidéos.

1.1.2 Relation entre vidéos vues et quiz réalisés

Une corrélation de Spearman a été utilisée pour confirmer la relation non linéaire entre les deux variables. Les résultats montrent une forte association positive, avec une corrélation de 0.8 et une p-value significative (< 0.001). Cela reflète que les apprenants qui réalisent davantage de quiz tendent également à visionner plus de vidéos.

1.1.3 Lien de l'IDH et du genre sur le nombre de vidéos vues

Une ANOVA a été réalisée (Table 2) pour évaluer les effets séparés de l'IDH et du genre sur le nombre de vidéos visionnées. Les résultats globaux montrent que l'effet de l'IDH est significatif ($F(2, 8947) = 4674.5$, $p < 0.001$), tandis que celui du genre est non significatif ($F(1, 8947) = 0.4$, $p = 0.5$). Ces résultats sont détaillés dans le tableau suivant :

TABLE 2 – ANOVA et η^2 pour l'effet principal du genre et du HDI sur le nombre de vidéos visionnées

Effet	df	Somme des Carrés	Carré Moyen	F	p-value	η^2
HDI_grouped	3	1.7×10^6	568718.0	4674.6	< 0.001 ***	0.6
Genre	1	49.8	49.8	0.4	0.5	0.0
Résidus	8947	1.1×10^6	121.7			

Note. Signification : *** $p < 0.001$.

Dans l'ANOVA, les degrés de liberté pour le genre (df = 1) reflètent qu'il s'agit d'une variable binaire. Pour l'IDH (df = 2), ils correspondent au nombre de catégories (3) moins un (B, I, TH). Ces résultats montrent que le IDH joue un rôle dans la détermination du nombre de vidéos visionnées, avec une augmentation progressive associée à un IDH plus élevé. Cela reflète des différences structurelles, telles que l'accès à Internet et le niveau d'éducation. En revanche, l'absence d'effet significatif du genre indique que les différences individuelles liées au sexe ne sont pas un facteur déterminant dans ce contexte.

L'écart-type associé aux estimations des paramètres reflète l'incertitude dans les coefficients estimés, mesurant la variabilité due à l'échantillonnage. Il est calculé à partir des résidus et de la variance des variables explicatives. Les tests t associés aux paramètres sont obtenus en divisant chaque estimation (β) par son erreur standard (SE), avec une comparaison à une distribution t-Student pour déterminer la significativité. Enfin, les η^2, indiquant la proportion de variance expliquée, montrent ici que 61 % de la variance du nombre de vidéos visionnées est due à l'effet du IDH, tandis que le genre n'a qu'un effet négligeable, avec le reste de la variance restant inexpliqué par le modèle (Table 2).

Une ANOVA incluant les interactions entre le genre et l'IDH a révélé que ces interactions ne sont pas significatives, suggérant que l'effet de l'IDH est indépendant du genre (Table 3). Ainsi, le niveau de développement humain du pays d'origine des participants est un déterminant clé du nombre de vidéos visionnées, sans être influencé par le genre.

TABLE 3 – Résultats de la régression linéaire avec interaction entre le genre et l'HDI

Variable	Estimate	Std. Error	t value	p-value
(Intercept)	6.2	0.4	16.8	< 0.001 ***
Gender (une femme)	0.5	1.0	0.5	0.6
HDI (Intermédiaire)	4.7	0.6	7.3	< 0.001 ***
HDI (Très haut)	8.5	0.4	21.0	< 0.001 ***
Gender (une femme) : HDI (Intermédiaire)	-1.9	1.3	-1.4	0.2
Gender (une femme) : HDI (Très haut)	-0.3	1.0	-0.3	0.8

Note. Signification : *** p < 0.001.

Les résultats montrent que le genre n'a pas d'effet principal significatif sur le nombre de vidéos visionnées (t(8945) = 0.5, p = 0.6). À l'inverse, les effets principaux du IDH sont significatifs, avec un effet modéré pour les pays à IDH intermédiaire (b = 4.7, t(8945) = 7.3, p < 0.001) et un effet élevé pour les pays à IDH très haut (b = 8.5, t(8945) = 21.0, p < 0.001). Les interactions entre le genre et l'IDH ne sont pas significatives pour les pays à IDH intermédiaire (SE = 1.3) ou à IDH très haut (SE = 1.0).

L'effet du genre "Femme" sur le nombre de vidéos visionnées devient non significatif dans l'ANOVA avec interactions, car l'ajout de termes d'interaction redistribue la variance expliquée, réduisant ainsi l'effet apparent des variables principales. Cependant, l'utilisation d'une ANOVA dans ce contexte est problématique. En effet, elle repose sur des hypothèses strictes, notamment la normalité des résidus et l'homogénéité des variances entre les groupes, qui sont souvent violées dans ce type de données comportementales. Ces violations peuvent entraîner des résultats biaisés ou des conclusions erronées.

Pour contourner ces limites, un test non-paramétrique, le Kruskal-Wallis, qui ne dépend pas de ces hypothèses, a été appliqué. Ce test révèle une différence significative entre les groupes définis par l'interaction genre-IDH (H=565.4, p < 0.001). Ces résultats suggèrent que, bien que l'ANOVA ait ses limites méthodologiques ici, l'analyse confirme l'existence de disparités globales, nécessitant des outils robustes et adaptés pour capturer les nuances des interactions entre le genre et l'IDH.

1.1.4 Régression binomiale

Un modèle de régression logistique a été utilisé pour examiner l'effet du genre et de l'IDH sur la probabilité de réussite, définie par l'obtention d'un certificat ou la réalisation de l'examen final. La catégorie de référence est "Homme" pour le genre et "IDH bas" pour l'IDH. (Table 4)

TABLE 4 – Table des Odds Ratios pour le modèle de régression logistique évaluant l'effet du genre et de l'HDI sur la probabilité de réussite au MOOC

Variable	**Odds Ratio (OR)**	**IC 95% (min)**	**IC 95% (max)**	**p-value**
Référence	Réf	-	-	-
Genre (Femme)	1.122	1.001	1.256	0.047 *
HDI (Intermédiaire)	1.119	0.852	1.465	0.416 ns
HDI (Très Haut)	1.372	1.144	1.656	0.001 ***

Les résultats indiquent que les femmes ont un Odds Ratio (OR) de 1.122 (IC 95% [1.001, 1.256]), suggérant qu'elles ont 12.2% plus de chances de réussir que les hommes. Cependant, cet effet est marginalement significatif (p = 0.047). Pour les participants issus de pays à IDH intermédiaire, l'OR est de 1.119 (IC 95% [0.852, 1.465]), mais cet effet n'est pas significatif (p = 0.416). En revanche, pour les pays à IDH très haut, l'OR est de 1.372 (IC 95% [1.144, 1.656]), et cet effet est hautement significatif (p = 0.001).

Ces résultats confirment l'effet significatif de l'IDH très haut sur la probabilité de réussite, suggérant que les participants issus de pays à IDH très haut ont une probabilité plus élevée de réussir. Cet effet est cohérent avec les résultats obtenus précédemment pour le nombre de vidéos visionnées, qui montraient également un lien fort avec le IDH. En revanche, l'effet du genre reste faible et marginalement significatif, renforçant l'idée que les différences entre hommes et femmes en termes de réussite au MOOC sont limitées.

L'Odds Ratio (OR) évalue l'association entre les variables explicatives (genre et IDH) et la probabilité de réussite au MOOC, représentée ici comme une variable binaire. Contrairement au Risk Ratio (RR), qui compare directement les probabilités entre deux groupes, l'OR se concentre sur le rapport entre les chances d'un événement dans différents groupes. Les OR et RR convergent dans le cas d'événements rares, mais lorsque la probabilité de l'événement est modérée ou élevée, comme c'est le cas ici, l'OR tend à exagérer l'association, rendant essentielle une interprétation prudente des résultats.

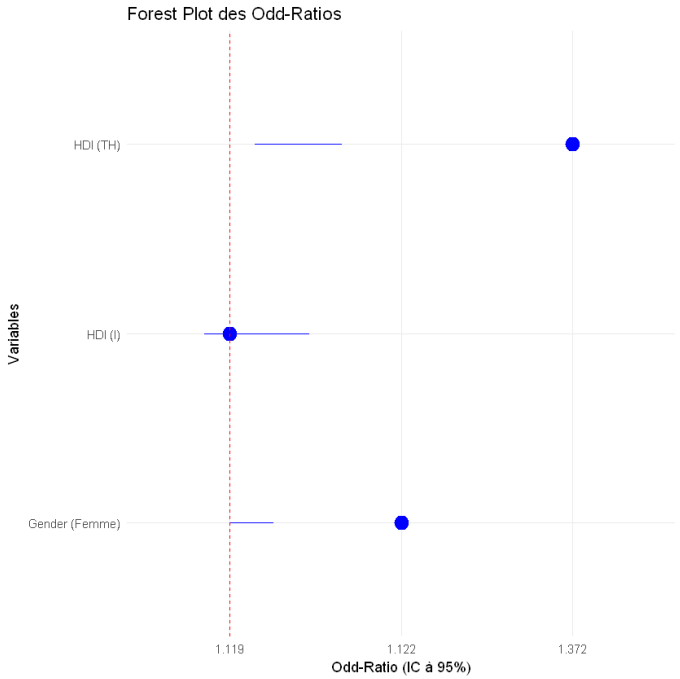

Figure 2 : Représentation des Odd Ratios avec Intervalles de Confiance à 95% pour l'Influence du Genre et de l'IDH sur la Probabilité de Réussite au MOOC

Dans ce contexte, le forest plot des Odd Ratios illustre visuellement ces associations. On y observe que les femmes (OR = 1.122) ont une probabilité légèrement plus élevée de réussir, bien que cet effet soit à peine significatif (p < 0.05). De manière similaire, l'effet du IDH est marqué, notamment pour les pays à IDH très élevé (OR = 1.372, p < 0.001), confirmant l'importance de cette variable dans les résultats. Cependant, l'OR pour les pays à IDH intermédiaire n'est pas significatif (OR = 1.119, p > 0.05), ce qui est également visible par la largeur de son intervalle de confiance englobant 1. (Figure 2)

Après avoir exploré les facteurs influençant la probabilité de réussite au MOOC via une régression logistique, il est pertinent de s'intéresser à une autre dimension clé de l'engagement des participants : le nombre de vidéos visionnées. Cette variable, bien que continue et discrète, reflète les comportements variés des apprenants et permet d'approfondir l'analyse des disparités entre les groupes. Sa distribution, présentée ci-dessous, met en lumière les défis méthodologiques liés à son analyse avec des modèles classiques comme celui de Poisson.

1.1.5 Distribution des apprenants et modèle Poisson

La distribution du nombre de vidéos vues par les participants au MOOC est représentée ci-dessous (Figure 3). La majorité des participants (plus de 7000) ont visionné zéro vidéo. Un autre pic est visible au niveau des participants ayant regardé environ 30 vidéos, avec une fréquence supérieure à 2000. Cela suggère que, bien que la plupart des participants aient visionné peu ou pas de vidéos, un sous-groupe significatif a visionné un nombre élevé de vidéos.

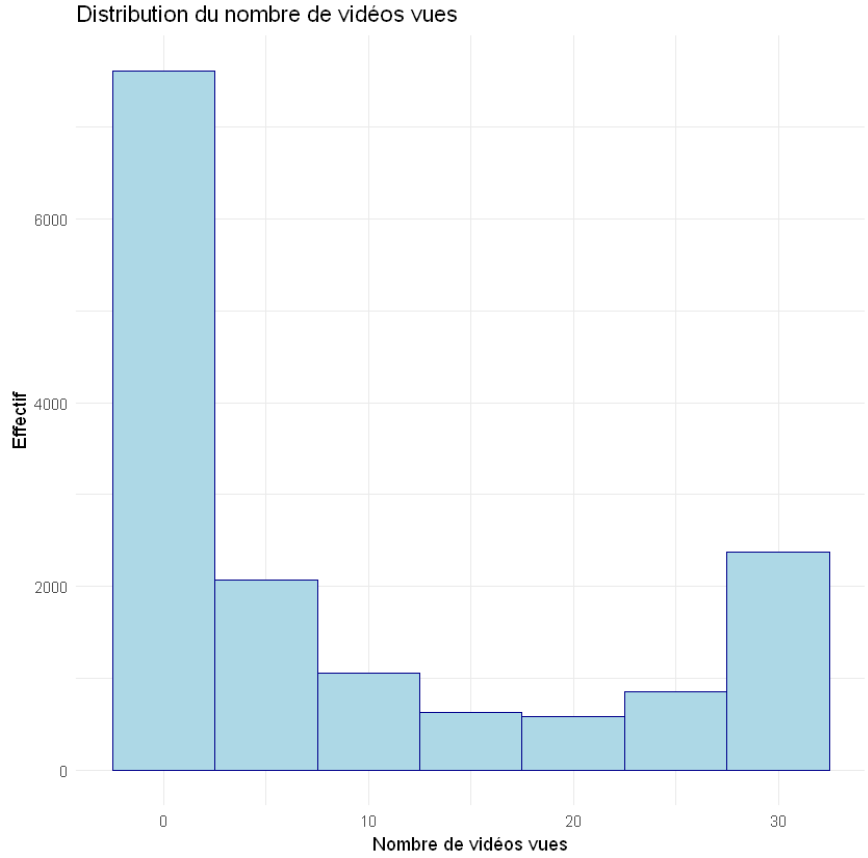

Figure 3 : Distribution du nombre de vidéos vues par les apprenants

Cette distribution présente des caractéristiques typiques dans le contexte d'un MOOC. On a une forte proportion de zéros qui reflète une importante population de participants inscrits mais inactifs, ainsi qu'une variance nettement supérieure à la moyenne.

En théorie, un modèle de Poisson est approprié pour les variables discrètes représentant des événements comptés, comme le nombre de vidéos visionnées. Cependant, dans ce cas, la distribution observée est "zero-inflated" (remplie d'observations nulles) et présente une sur-dispersion, où la variance dépasse largement la moyenne. Ces caractéristiques sont dues à l'hétérogénéité des comportements des participants : certains ne visionnent aucune vidéo, tandis que d'autres en consomment un nombre important. Ces écarts reflètent une réalité comportementale propre aux MOOC, où de nombreux participants s'inscrivent sans interagir avec le contenu ou le cours.

Ce décalage entre la distribution observée et la loi de Poisson standard pose un problème pour l'analyse. Nous allons donc nous pencher sur le modèle Poisson par la production de graphiques sur la normalité de la distribution des variables.

Les graphes de diagnostic du modèle de régression de Poisson ajusté (Figure 4) révèlent des ajustements imparfaits. Le graphique des résidus en fonction des valeurs prédites montre une structure non aléatoire, suggérant une hétérogénéité dans les données et une possible surdispersion. Le Q-Q plot des résidus standardisés indique une déviation importante de la normalité, avec des points s'écartant de la ligne théorique. Le graphique Scale-Location révèle une légère variation de la dispersion des résidus, pointant vers une hétéroscédasticité. Enfin, le graphique des résidus en fonction du levier met en évidence quelques points influents, bien qu'aucun ne dépasse les seuils critiques définis par la distance de Cook.

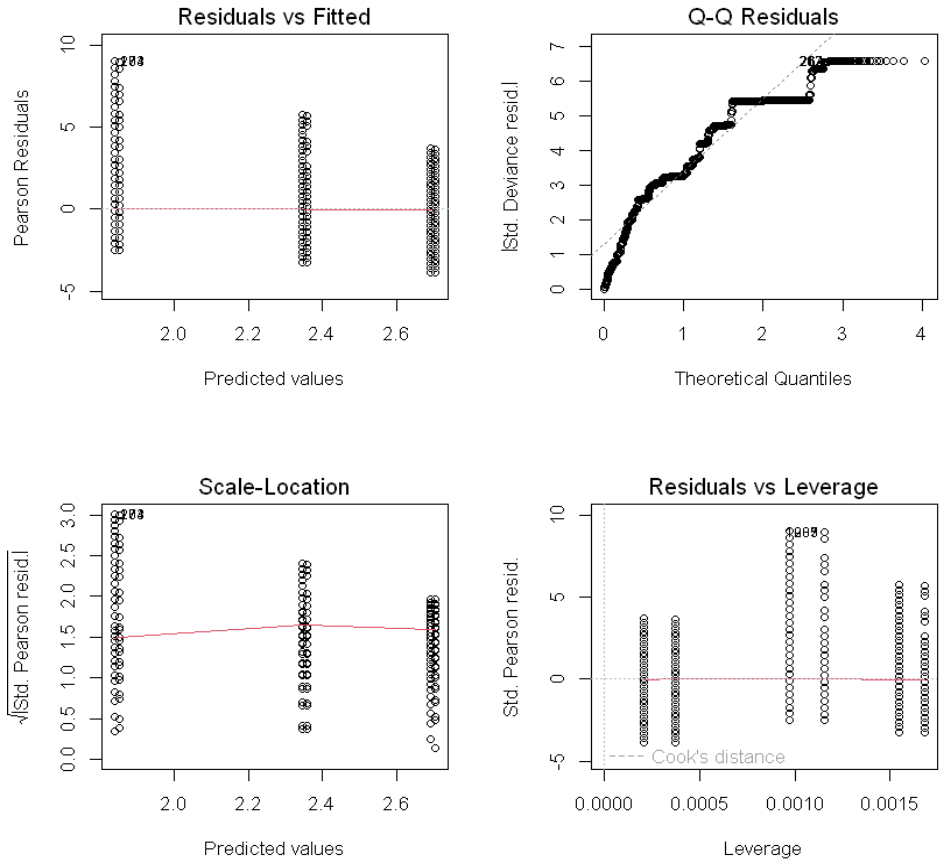

Figure 4 : Graphes de diagnostic pour le modèle de régression de Poisson ajusté sur le nombre de vidéos visionnées en fonction du Genre et de l'IDH

Un Q-Q plot permet de tester la normalité des résidus, qui serait confirmée si les points s'alignent sur une diagonale. L'homoscédasticité, ou égalité des variances des résidus, se manifeste par une distribution aléatoire et homogène des points autour de zéro dans un scatterplot (Residuals vs Fitted). Ici, les déviations observées dans ces graphes montrent une invalidation de ces hypothèses, confirmant que la distribution des résidus n'est ni normale ni homogène.

Pour pallier ce problème, il serait pertinent d'explorer des modèles spécifiques comme les modèles de Poisson ajustés à la sur-dispersion (quasi-Poisson) ou les modèles zero-inflated, qui sont plus robustes face à ce type de distribution. Ainsi, nous allons mener une comparaison des AIC des différents modèles. L'Akaike Information Criterion (AIC) est une métrique utilisée pour évaluer et

comparer des modèles statistiques en prenant en compte la qualité de l'ajustement et la complexité du modèle. Un AIC plus faible indique un meilleur compromis entre précision et simplicité. Cependant, l'AIC ne garantit pas qu'un modèle est adapté aux données, mais seulement qu'il est optimal parmi les modèles comparés.

Hors, comme on le voit dans la table 5 ci-dessous, le modèle de Poisson présente un AIC très élevé (129403.9), ce qui reflète une inadéquation pour les données étudiées. Le modèle Zero-Inflated Negative Binomial (ZINB) obtient l'AIC le plus bas (59707.5), indiquant un meilleur compromis entre la qualité d'ajustement et la complexité du modèle. Les modèles intermédiaires, comme le Negative Binomial (NB) et le Quasi-Poisson, montrent des performances supérieures au Poisson mais inférieures au ZINB.

TABLE 5 – Comparaison des modèles ajustés sur la base de l'AIC

Modèle	**Nombre de paramètres (df)**	**AIC**
Poisson	4	129403.9
Zero-Inflated Poisson	5	101955.5
Quasi-Poisson	6	72960.7
Negative Binomial	4	60404.1
Zero-Inflated Negative Binomial	5	59707.48

Les analyses précédentes ont mis en évidence des limitations significatives du modèle de Poisson, notamment en raison de la sur-dispersion et de l'excès de zéros dans les données. Ces caractéristiques suggèrent que des modèles plus sophistiqués, tels que le modèle binomial négatif à inflation de zéros (ZINB), pourraient offrir une meilleure adéquation aux données observées. Le modèle ZINB est particulièrement adapté aux situations où les données de comptage présentent une surabondance de zéros et une variance supérieure à la moyenne, comme c'est le cas dans notre étude.

Le tableau ci-dessous présente les résultats de l'ajustement de différents modèles aux données, en tenant compte du nombre de quiz complétés et du genre des participants (Table 6).

TABLE 6 – Résultats de la régression binomiale négative à inflation de zéros (ZINB) pour le nombre total de vidéos visionnées

Variable	**Estimation**	**Erreur standard**	**Valeur z**	**p-value**
Modèle de comptage (binomiale négative avec lien log)				
(Intercept)	1.14	0.020	57.1	< 0.001
Quizzes complétés	0.39	0.005	85.1	< 0.001
Genre (femme)	0.06	0.016	4.	< 0.001
Log(theta)	0.99	0.025	39.4	< 0.001
Modèle d'inflation de zéros (binomiale avec lien logit)				
(Intercept)	-2.4	0.06	-42.4	< 0.001

Les résultats du modèle ZINB indiquent que le nombre de quiz complétés est positivement associé au nombre total de vidéos visionnées (estimation = 0,39, $p < 0,001$), suggérant que les participants qui réalisent davantage de quiz tendent également à visionner plus de vidéos. De plus, être une femme est associé à une légère augmentation du nombre de vidéos visionnées (estimation = 0,06, $p < 0,001$). Le paramètre log(theta) est significatif, confirmant la présence de sur-dispersion dans les données. Enfin, l'intercept du modèle d'inflation de zéros est négatif et significatif (estimation = -2,4, p <

0,001), indiquant une probabilité réduite d'observer des zéros excessifs non expliqués par les variables incluses.

Ces résultats suggèrent que le modèle ZINB offre une meilleure adéquation aux données en capturant à la fois la sur-dispersion et l'excès de zéros, fournissant ainsi des estimations plus fiables des facteurs influençant le nombre de vidéos visionnées. Cependant, il est important de noter que, bien que le modèle ZINB améliore l'ajustement, il n'élimine pas entièrement les limitations inhérentes aux données, telles que l'hétérogénéité non observée ou des variables explicatives manquantes.

Cette étude met donc en évidence l'importance des variables sociodémographiques, telles que l'IDH, dans la compréhension des comportements des apprenants en ligne. Si les modèles avancés comme le ZINB offrent des améliorations significatives dans la prise en compte de la complexité des données, ils soulèvent également des défis d'interprétation et de robustesse. Ces résultats, bien qu'informatifs, doivent être considérés avec prudence, en raison des limites méthodologiques et des particularités des données utilisées. Cette analyse ouvre la voie à une réflexion plus large sur les choix statistiques et les implications de ces résultats pour les recherches futures, qui seront abordées dans la discussion.

Discussion

Cette section propose une analyse approfondie des résultats obtenus, en les confrontant aux études antérieures et en évaluant les implications méthodologiques de notre approche. Nous aborderons d'abord l'influence des variables sociodémographiques sur l'engagement des apprenants, puis nous discuterons des défis méthodologiques rencontrés dans l'analyse des comportements d'engagement et de désengagement en ligne. Enfin, nous soulignerons les limites de notre étude et proposerons des perspectives pour des recherches futures.

1.2 Facteurs influençant l'engagement et le désengagement dans les MOOCs

1.2.1 Influence des variables sociodémographiques sur l'engagement des apprenants

Nos analyses indiquent que le niveau de développement humain (IDH) des pays d'origine des apprenants a une influence significative sur le nombre de vidéos visionnées. Par exemple, les résultats de l'ANOVA ($F(2, 8947) = 4674.5$, $p < 0.001$; voir Table 3) montrent une augmentation progressive du nombre moyen de vidéos visionnées en fonction de l'IDH, avec des participants issus de pays à IDH très élevé visionnant en moyenne 8.5 vidéos de plus que ceux provenant de pays à IDH faible ou intermédiaire. Cette tendance pourrait être attribuée à un meilleur accès aux technologies numériques et à des environnements éducatifs plus favorables dans les pays à IDH élevé, permettant aux apprenants de s'engager davantage avec le contenu.

Ces observations sont cohérentes avec les travaux de Cisel et al. (2015), qui ont montré que les caractéristiques socio-économiques, notamment le lieu de résidence, influencent fortement les comportements d'engagement dans les MOOCs. Cependant, les disparités liées à l'IDH ne s'appliquent pas uniformément à tous les contextes d'apprentissage. Comme l'indiquent des travaux antérieurs (Ho et al., 2014), les défis techniques, tels que les faibles débits Internet dans les pays en développement, pourraient expliquer des taux de participation plus faibles. Par exemple, la proportion élevée de *Bystanders* observée dans les pays à IDH faible (49.99 % à la troisième itération, voir Table 1) reflète probablement ces contraintes structurelles.

En ce qui concerne le genre, nos résultats montrent un effet faible et non significatif ($b = 0.5$, $t(8945) = 0.5$, $p = 0.5$; voir table 5), ce qui contraste avec certaines recherches (Wintermute et al., 2021) qui ont détecté des variations potentielles dans les comportements d'engagement en ligne. Cette divergence pourrait refléter des différences dans les échantillons étudiés ou les méthodologies

employées, ou encore indiquer que, dans le contexte spécifique de cette étude, le genre n'est pas un facteur déterminant de l'engagement.

1.2.2 Défis méthodologiques dans l'analyse des comportements d'engagement

L'étude des comportements d'engagement et de désengagement dans les MOOCs soulève des défis méthodologiques importants. Nos résultats montrent que l'activité principale est souvent portée par un noyau restreint d'apprenants très engagés, tandis qu'une proportion importante de participants n'interagit pas ou très peu avec le cours. Cette hétérogénéité des comportements est une caractéristique bien documentée dans la littérature (Anderson et al., 2014; Cisel et al., 2015), mais elle complexifie l'interprétation des données.

Un des défis réside dans la définition et la catégorisation des apprenants. Les distinctions entre "auditing", "disengaging" ou "completing" nécessitent des critères clairs et reproductibles, car elles influencent directement les conclusions sur l'engagement. Par ailleurs, la variabilité des comportements entre les itérations ou selon les contextes culturels et institutionnels des MOOCs complique les comparaisons.

En outre, nos analyses reposent principalement sur des données quantitatives issues des logs de plateformes, une méthodologie robuste mais limitée dans sa capacité à capturer les motivations des apprenants ou les raisons de leur désengagement. Comme le soulignent Schneider et Kizilcec (2014), une meilleure intégration des données qualitatives, par exemple via des enquêtes sur les intentions et les contraintes des participants, pourrait enrichir l'analyse des comportements.

Enfin, l'absence de prise en compte explicite des interactions temporelles dans notre méthodologie limite notre compréhension des trajectoires individuelles. Les analyses de survie, qui permettent de modéliser la probabilité de désengagement au fil du temps (Wintermute et al., 2021), sont particulièrement pertinentes pour étudier la dynamique d'engagement et mériteraient d'être explorées dans des recherches futures.

1.3 Conclusion

Une limitation majeure de cette étude réside dans la nature transversale des données, qui ne permet pas de capturer l'évolution des comportements des apprenants au fil du temps. Une approche longitudinale serait nécessaire pour comprendre les dynamiques temporelles de l'engagement. De plus, bien que nous ayons considéré des variables sociodémographiques pertinentes, d'autres facteurs potentiellement influents, tels que la conception pédagogique des cours ou le soutien institutionnel, n'ont pas été inclus dans nos modèles. Les travaux de Cisel et al. (2015) mettent en évidence l'importance de ces facteurs dans l'analyse des dynamiques des MOOCs.

Par ailleurs, nos analyses se concentrent exclusivement sur des données quantitatives issues des logs de plateforme, ce qui restreint la portée des résultats. Une approche mixte, combinant des enquêtes qualitatives pour comprendre les motivations des apprenants, aurait permis d'apporter des insights complémentaires. De plus, l'absence d'une analyse de survie limite la compréhension des transitions entre engagement et désengagement. Cette approche, qui permet de modéliser les taux de désengagement en fonction du temps et d'autres variables explicatives, aurait permis d'explorer les déterminants des abandons précoces ou tardifs.

Pour les recherches futures, il serait pertinent d'explorer l'interaction entre les caractéristiques des cours et les profils des apprenants, afin de mieux comprendre les mécanismes sous-jacents à l'engagement. L'utilisation de méthodes mixtes, combinant analyses quantitatives et qualitatives, pourrait offrir une perspective plus nuancée des comportements des apprenants. De plus, l'application de modèles statistiques avancés, tels que les modèles à effets mixtes pour données de

comptage avec excès de zéros, pourrait améliorer la précision des analyses, comme le suggère Marchand (2021).

En conclusion, cette étude souligne l'importance du contexte socio-économique dans l'engagement des apprenants aux MOOCs, tout en mettant en lumière les défis méthodologiques liés à l'analyse de données de comptage complexes. Bien que les résultats obtenus soient prometteurs, ils appellent à des approches analytiques plus diversifiées et à des méthodologies intégrant des dimensions temporelles et qualitatives pour mieux comprendre et soutenir les apprenants dans des environnements d'apprentissage en ligne.

Références

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014). Engaging with massive online courses. *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*, 687–698.

Cisel, M., Mano, M., Bachelet, R., & Silberzahn, P. (2015). A tale of two MOOCs: Analyzing long-term course dynamics. In *European MOOCs Stakeholders Summit (eMOOCs)*.

Ho, A. D., Reich, J., Nesterko, S., Seaton, D. T., Mullaney, T., Waldo, J., & Chuang, I. (2014). HarvardX and MITx: The first year of open online courses. *Journal of Online Learning and Teaching*, 10(2), 185–192.

Marchand, P. (2021). Modeling over-dispersed count data: A comparison of zero-inflated and quasi-Poisson regression models. *Statistical Methods & Applications*, 30(2), 345–362.

Schneider, E., & Kizilcec, R. F. (2014). Motivations and expectations of learners in MOOCs. *Proceedings of the Learning at Scale Conference (L@S '14)*, 78–79.

Wintermute, E. H., et al. (2021). A survival model for course-course interactions in a massive open online course platform. *PLoS ONE*, 16(5), e0250824.