Réplication d'un article : « L'université française, lieu de brassage ou de ségrégation sociale ? »

Projet d'économétrie — Département de Sciences Humaines et Sociales

École normale supérieure Paris-Saclay

Jean-Baptiste Lagrange-Dupuis

20 décembre 2023

Table des matières

Liste des tableaux

Table des figures

Toutes nos analyses ont été réalisées à l'aide du logiciel statistique R (version 4.2.1, R Core Team 2022). Les différents paquets utilisés sont listés en page 13. Ce document a été écrit avec LaTeX.

Introduction

Dans la suite de ce travail nous nous proposons de répliquer un article [Avouac et Harari-Kermadec 2022]. Après avoir résumé l'article, nous présenterons brièvement la base de données utilisée ainsi que les transformations mises en œuvre, puis nous examinerons deux modèles, l'un linéaire et l'autre Tobit, visant à expliquer respectivement le financement des universités et le classement de Shanghai.

1 Résumé de l'article

Dans leur article datant de 2022, Romain Avouac et Hugo Harari-Kermadec se proposent d'étudier les transformations récentes (2007-2015) des universités françaises, en particulier la question de leur polarisation. À partir des données du Système d'information sur le suivi de l'étudiant (SISE), ils procèdent en trois étapes : d'abord la mise en évidence de l'hétérogénéité de l'espace universitaire grâce à une ACP, puis des mesures de ségrégation et de polarisation visant à rendre compte de cette hétérogénéité et enfin la mise en relation de la polarisation constatée avec les dispositifs qui modifient la structure du champ universitaire, à l'instar des classements universitaires et des initiatives d'excellence.

Les chercheurs partent du constat que les classements internationaux, reflets de la globalisation de l'enseignement supérieur, ont conduit à une volonté politique de créer des établissements de renommée internationale qui bénéficient de dotations plus importantes, ce dont témoignent les « initiatives d'excellence » (Idex). Or l'une des conséquences de ces nouvelles modalités d'attribution et de répartition des financements au sein de l'Université française pourrait être une polarisation des établissements, avec un système éducatif « à deux vitesses », où les universités de second rang (par opposition aux universités de rang mondial) se partageraient la portion congrue.

Après un retour sur la massification de l'accès au supérieur en France ces dernières décennies, avec une diversification du profil et de l'origine des étudiant·es, les auteurs annoncent leur problématique : « Observe-t-on effectivement une polarisation du système universitaire français en matière de composition sociale ? ». Dans la suite du résumé, nous reprendrons le plan en trois parties de l'article.

1.1 Un paysage universitaire socialement hétérogène

Les auteurs mènent une Analyse en Composantes Principales (ACP) afin de visualiser la distribution des caractéristiques sociales et scolaires des populations étudiantes, en s'appuyant sur des variables telles que la catégorie socio-professionnelle du premier parent référent (qui caractérise ici l'origine sociale), la série du bac, le sexe des étudiant·es ainsi que leur cycle d'inscription (le statut de boursier de supérieur et l'échelon associé sont laissés de coté car ces variables ne sont pas disponibles pour toutes les périodes).

L'ACP, dont on a retenu les deux premières composantes, ordonne horizontalement les établissements selon leur public, d'origine plus ou moins favorisée. Les variables qui contribuent le plus à cet ordonnancement ont trait à la profession des parents (CPIS, ouvrier non qualifiés ou employés) et au bac (professionnel ou technologique). Le second axe distingue les établissements à dominante littéraire de ceux à composante scientifique, y contribuent le plus les variables correspondant aux séries du bac. Lorsque l'ACP porte sur les établissements, elle montre que leurs publics sont très hétérogènes socialement, les enfants de CPIS (respectivement d'ouvriers) étant surreprésentés dans les universités parisiennes (respectivement du Nord et Sud-Est).

À partir de ces constatations, les auteurs établissent une typologie des établissements français en réalisant une classification ascendante hiérarchique à partir des deux dimensions principales de l'ACP, qui les conduit à distinguer quatre groupes, selon deux critères de sélection détaillés ci-après :

— un premier critère distingue les établissements accueillant des publics d'origine sociale intermédiaire, ceux accueillant des publics populaires et ceux à composition sociale favorisée ;

— un second critère établit deux catégories, au sein des établissements à composition sociale favorisée, selon que ces établissements sont à dominante littéraire ou scientifique, l'hétérogénéité pour ce critère étant plus marquée au sein des établissements accueillant des publics favorisés.

Cette polarisation plus importante au sein des établissements accueillant des publics favorisés s'explique aussi par l'effet des baccaularéats scientifiques, ces derniers étant prépondérants dans les établissements favorisés à dominante scientifique, là où les établissements littéraires accueillent des publics issus de bacs ES/L, de même que les établissements fréquentés par des étudiants issus de classes plus populaires, où sont présents davantage de bacheliers profesionnels et techniques.

1.2 Une polarisation sociale croissante des établissements d'enseignement supérieur français

Cette seconde partie vise à tester deux hypothèses : d'une part l'hypothèse d'un mouvement de polarisation de l'enseignement supérieur français et d'autre part l'émergence d'une dimension ségrégative à la suite de la démocratisation de l'accès à l'université.

Pour rendre compte de la réalité de cette polarisation, les auteurs se proposent de mesurer l'évolution du barycentre de chacun des groupes d'établissements, et observent que si le groupe des compositions sociales intermédiaires et celui des compositions sociales favorisées à dominante littéraire tendent à s'ouvrir légèrement, les deux groupes restants (groupe à composition sociale favorisée et dominante scientifique et groupe accueillant des publics populaires) tendent à se fermer.

Les auteurs conduisent d'abord une analyse visant à rendre compte de la ségrégation dans l'enseignement supérieur en utilisant l'indice d'entropie normalisé, qui mesure comment les catégories sociales du parent référent se répartissent dans les différents groupes de la typologie établie précédemment, et dont la baisse indique une réduction de la ségrégation.

Il s'agit ensuite de tester l'hypothèse d'une polarisation des établissements d'enseignement supérieur français relativement à leur composition sociale. Les auteurs choisissent d'appliquer le cadre économétrique de la polarisation [Esteban et Ray 1994], dont il ressort un accroissement de la polarisation sociale des établissements supérieurs français entre 2005 et 2017, de l'ordre de 20% (pour la typologie à quatre groupes).

Deux raisons sont avancées pour expliquer ces évolutions divergentes des mesures de ségrégation et de polarisation. D'une part, la baisse de l'indice d'entropie peut s'expliquer par un reforcement de l'homogénéité de chaque catégorie de la typologie prise isolément plutôt que par un rapprochement des barycentres des différentes catégories. D'autre part, là où l'indice d'entropie n'est calculé qu'à partir de la catégorie sociale du parent référent, l'analyse de la polarisation prend en compte davantage de critères afin de définir l'origine sociale des étudiant·es, ce qui peut expliquer l'évolution divergente de ces indicateurs : les auteurs constatent en effet une très forte augmentation du nombre de titulaires d'un baccalauréat professionnel dans l'enseignement supérieur, augmentation captée essentiellement par les établissements accueillant des publics populaires, où les baccalauréats professionnels étaient déjà nombreux, ce qui explique en partie la dynamique de polarisation sur la période 2007-2015.

1.3 Les « politiques d'excellence » de l'enseignement supérieur : une redistribution inversée ?

Dans cette troisième partie, les auteurs se donnent pour objectif d'identifier les causes du processus d'accroissement de l'hétérogénéité sociale des établissements entre 2007 et 2015. Ils envisagent d'abord la possibilité que cette polarisation soit liée à une augmentation des inégalités sociales entre les territoires, hypothèse qui est vite écartée du fait d'une forte propension à la mobilité géographique chez les étudiant·es. Ceci les conduit à étudier le lien entre la composition sociale des établissements et le fait d'être distingué par des dispositifs d'excellence, qui orientent en partie les politiques récentes de l'enseignement supérieur et constituent une source importante de revenus pour les établissements. À partir d'une ACP, les auteurs montrent que les universités labellisées IDEX accueillent des populations étudiantes plutôt favorisées tandis que les I-SITE correspondent à des universités à composition sociale intermédiaire.

Les auteurs concluent leur article par le constat d'une forte hétérogéneité sociale du système universitaire français, également caractérisé par une faible mixité sociale au sein des établissements.

2 Présentation de la base de données

Les analyses qui suivent sont réalisées à partir des données de la base de données du SISE (Système d'information sur le suivi de l'étudiant) et de données sur le financement des universités provenant du package kpiESR de Julien Gossa. Ces deux bases de données ont été reliées après que les données du SISE, dont chaque observation correspond à un·e étudiant·e, ont été agrégées par établissement de façon à obtenir des moyennes des variables pertinentes pour chacun des établissements.

De nouvelles variables composites ont été crées, à commencer par la variable `cohorte_shanghai` qui est un recodage du classement de shanghai : on a créé six catégories d'établissements à partir de leur rang au classement 2021 (lorsqu'on ne dispose pas du rang 2021, on prend celui de 2019 s'il existe, sinon on considère que le rang de l'établissement est strictement supérieur à 1000) :

$$\texttt{cohorte\_shanghai} = \begin{cases} 0 \text{ si } \texttt{class\_shanghai} \leq 20 \\ 1 \text{ si } \texttt{class\_shanghai} \leq 50 \\ 2 \text{ si } \texttt{class\_shanghai} \leq 100 \\ 3 \text{ si } \texttt{class\_shanghai} \leq 200 \\ 4 \text{ si } \texttt{class\_shanghai} \leq 500 \\ 5 \text{ si } \texttt{class\_shanghai} \leq 1000 \\ 6 \text{ si } \texttt{class\_shanghai} > 1000 \end{cases}$$

Une variable `Init_Excel` a également été créée pour rendre compte de la labellisation des établissements (labels I-SITE et IDEX). Plutôt que de créer deux variables correspondant respectivement à chacun de ces deux labels, nous avons décidé, étant donné la « hiérarchie » [1] entre ces labels de créer une unique variable codée comme suit :

$$\texttt{Init\_Excel} = \begin{cases} 0 \text{ si } \texttt{PerimEx} = \text{Non} \\ 1 \text{ si } \texttt{PerimEx} = \text{ISITE} \\ 2 \text{ si } \texttt{PerimEx} = \text{IDEx} \end{cases}$$

D'autres variables ont été créées pour connaître la proportion d'étudiant constituant des publics favorisés ou défavorisés pour chaque établissement. Pour ce faire, nous avons repris le groupement des PCS en quatre catégories utilisé par la DEPP [Rocher 2016], avec deux catégories favorisées, une catégorie moyenne et une catégorie défavorisée, en ne conservant qu'une catégorie favorisée, composée des étudiant·es ayant un parent référent cadre ou assimilé, chef d'entreprise, professeur des écoles ou assimilé, et une catégorie défavorisée, dont les étudiant·es ont un parent référent ouvrier ou inactif. Ces deux variables, qui correspondent donc à la proportion de parents appartenant à l'une ou à l'autre de ces catégories dans un établissement ont été nommées respectivement `public_favorisé` et `public_défavorisé`. Le tableau 3 présente la répartition de ces variables composites ainsi que des autres principales variables utilisées dans la suite de l'article.

3 Modèle linéaire pour expliquer le fonctionnement des universités

Nous cherchons à expliquer la subvention reçue par les établissements de la base de données à l'aide d'une régression linéaire. Nous considérons plusieurs potentielles variables explicatives des subventions : le nombre d'étudiants, le rang au classement de shanghai, la labellisation de l'établissement par une Initiative d'Excellence, et enfin le public accueilli par l'établissement, appréhendé par le biais du pourcentage d'étudiant·es défavorisé·es. Le tableau 1, qui présente les différents paramètres de la régression linéaire, montre que seules les variables `effectif` et `Init_Excel` sont significatives au seuil de 1%.

1. en effet, le gouvernement présente les I-SITE comme des universités « positionnées sur un nombre plus limité de thématiques » que les IDEX (`https://www.gouvernement.fr/idex-isite`)

TABLEAU 1 – Paramètres estimés de la régression linéaire

	Dependent variable :
	subvention
effectif	3 818,574***
	(596,581)
cohorte_shanghai	−1 262 368,000
	(8 566 585,000)
Init_Excel	95 237 706,000***
	(16 663 824,000)
public_défavorisé	85 453 977,000
	(107 695 710,000)
Constant	11 219 866,000
	(51 766 811,000)
Observations	82
R^2	0,758
Adjusted R^2	0,746
Residual Std. Error	68 071 170,000 (df = 77)
F Statistic	60,368*** (df = 4 ; 77)
Note :	*p<0,1 ; **p<0,05 ; ***p<0,01

Il est intéressant de noter que la variable relative au classement de shanghai n'est pas significative, ce qui ne nous permet donc pas de conclure à un lien entre le rang au classement de shanghai et le montant des subventions reçues par les établissements. De même, les types de publics accueillis par les établissements, appréhendés ici à travers le taux d'étudiant·es issu·es d'un milieu défavorisé, ne semble pas avoir d'influence sur le montant des subventions, alors que l'on aurait pu s'attendre à ce que certaines universités accueillant des publics plus populaires soient mieux dotés, où, à l'inverse à ce que les universités les mieux dotés soient des universités prestigieuses au recrutement sélectif où le taux d'étudiant·es issu·es de publics défavorisé est généralement plus faible, hypothèses que les résultats de la régression linéaire nous poussent à rejeter.

La variable `effectif` apparaît très significative, ce qui semble plutôt intuitif dans la mesure où le coût de fonctionnement des universités est fortement corrélé avec le nombre d'étudiant·es qui y sont inscrit·es. Mais plus encore que les effectifs, la variable la plus fortement significative est `Init_Excel`. Notons que la valeur très élevée du coefficient associé est à cette variable peut s'expliquer par le nombre très limité de modalités de cette variable (0, 1 ou 2), et que pour plus des trois quarts des universités, qui ne sont pas labellisées, cette variable vaut 0 (cf. Tableau 3 en annexe A). Néanmoins, au vu de ce coefficient, nous pouvons conclure que la labellisation des établissements par des initiatives d'excellence a un bien un effet sur les subventions reçues.

4 Modèle Tobit pour expliquer le classement de shanghai

À la lecture du tableau 2, on note que tous les coefficients du modèle sont différents de 0, bien que le coefficient de la variable `effectif` soit faible. Par ailleurs, les probabilités critiques des coefficients sont inférieures à 1%, sauf concernant les taux d'étudiant·es en doctorat et le pourcentage de femmes. Ces deux variables ne semblent donc pas significatives. Concernant les variables significatives, on observe que plus l'établissement compte d'étudiant·es, plus il est susceptible d'avoir un faible rang au classement de shanghai : en d'autres termes, les établissements les plus gros sont globalement mieux classés, ce qui est cohérent étant donné que le classement de shanghai s'appuie notamment sur le nombre de Prix Nobel et de médailles Fields parmi les diplômés et les professeurs notamment, ou sur le nombre de chercheurs parmi les plus cités dans leur discipline, ce qui tend à favoriser les plus grosses universités bien que ces valeurs soient rapportées au nombre d'étudiant·es et de professeurs que compte l'établissement.

La variable `Init_Excel` joue également négativement sur le classement : les établissements labellisés I-SITE ou, a fortiori, IDEX, sont mieux classés que les autres. C'est à nouveau plutôt intuitif dans la mesure où ces labels visent à identifier et à construire des universités de rang mondial. La question qui peut se poser est néanmoins celle de la causalité : la labellisation I-SITE ou IDEX est-elle à l'origine de moyens supplémentaires accordés aux universités concernées, ce qui leur permet de gagner des places dans le classement ? Ou vient-elle confirmer la position dominante des ces universités dans les versions précédentes du classement de shanghai ?

La variable `public_défavorisé` joue elle positivement sur le rang des universités au classement de Shanghai. Les universités qui accueillent davantage d'étudiant·es dont les parents appartiennent à des milieux moins favorisés, à la fois du point de vue des revenus, mais également du point de vue de la distance sociale par rapport à l'école, se voient moins bien classées. Ajoutons qu'il s'agit de la variable dont le coefficient est le plus important. Ce résultat vient à l'appui de travaux précédents [Avouac et Harari-Kermadec 2022] qui montraient que les universités présentes dans le classement de Shanghai accueillent souvent des publics plus favorisés.

Concernant la variable `master`, il est intéressant de voir qu'elle joue positivement sur le classement : les universités ayant une forte proportion d'étudiant·es en master sont moins bien classées. On aurait pu penser qu'à l'inverse, les universités ayant beaucoup d'étudiants en master et en doctorat étaient mieux classées. Mais une observation de la base de données montre qu'on trouve parmi les universités ayant les taux les plus importants d'étudiant·es en master majoritairement de petits établissements (notamment les IEP de Lille, Rennes et Lyon, l'Observatoire de Paris, les universités de technologie de Compiègne, Troyes et Belfort-Montbéliard, etc.), dont le rang au classement de Shanghai est nécessairement important.

TABLEAU 2 – Paramètres estimés du modèle probit

	Dependent variable :
	class_shanghai
effectif	$-0{,}021^{***}$
	(0,004)
Init_Excel	$-204{,}567^{***}$
	(76,476)
public_défavorisé	$6\ 594{,}593^{***}$
	(1 597,937)
master	$1\ 417{,}710^{***}$
	(421,068)
doctorat	286,383
	(638,393)
bac_S	$-1\ 772{,}406^{***}$
	(541,481)
femmes	$-523{,}799$
	(999,271)
Constant	860,991
	(823,615)
Observations	83
Log Likelihood	$-201{,}021$
Wald Test	$85{,}416^{***}$ (df = 7)
Note :	*p<0,1 ; **p<0,05 ; ***p<0,01

La très faible significativité du taux de doctorat peut s'expliquer par la très faible proportion de doctorats, quel que soit l'établissement : excepté pour quatre établissements (Comue Paris-Est et Bourgogne Franche-Comté, université PSL et INP Toulouse), le taux d'étudiant·es en doctorat est toujours inférieur à 1% .

Le taux de bacheliers scientifiques est corrélé négativement avec le classement, ce qui semble cohérent dans la mesure où le classement de shanghai est davantage centré sur les sciences fondamentales, notamment en ce qu'il prend en compte par exemple les articles publiés dans *Science* et *Nature*, ou encore le nombre de prix Nobel, dans la mesure où certaines disciplines comme la sociologie, la psychologie, les langues, ou encore la philosophie ne sont pas concernées par ce prix.

Enfin, la non significativité du taux de femmes indique une absence de lien entre cette variable et le classement de Shanghai, un résultat inverse aurait été étonnant.

Nos résultats concernant l'influence de la position sociale des parents des étudiant·es et du nombre d'étudiant·es sur le classement de shanghai sont congruents avec ceux de Romain Avouac [Avouac 2017], néanmoins contrairement à lui, nous trouvons une corrélation significative entre le fait d'être haut dans le classement pour un établissement et la part d'étudiant·es bacheliers scientifiques.

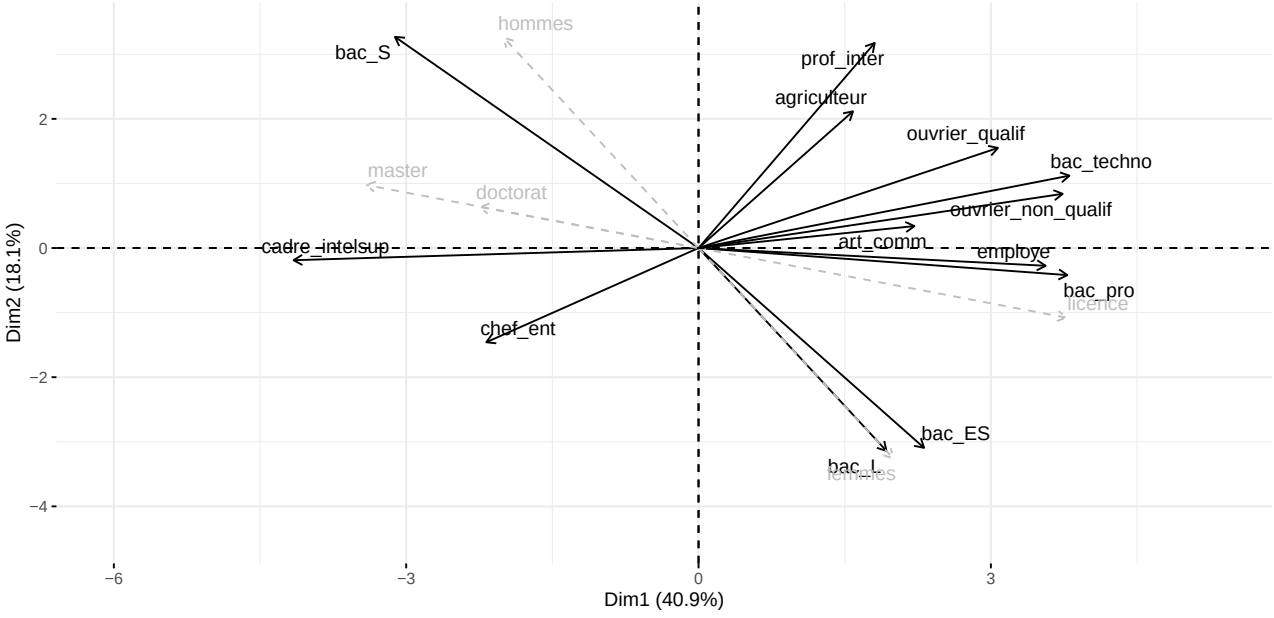
5 Analyse en composantes principales des universités françaises

Nous réalisons une analyse en composantes principales (ACP) sur les données SISE de 2007 que l'on compare à celle réalisée dans l'article de Romain Avouac et Hugo Harari-Kermadec [Avouac et Harari-Kermadec 2022], puis sur laquelle on projette les établissements tels qu'ils étaient en 2021 afin de mesurer les transformations sur la période.

5.1 Reproduction de l'ACP sur les données de 2007

Notons au sujet de cette ACP reproduite à partir de l'article mentionné plus haut que nous obtenons des chiffres assez similaires à ceux de Avouac et Harari-Kermadec 2022 pour la variance expliquée par chaque dimension, bien qu'inférieurs (41% et 19% respectivement pour nos premières et deuxième dimension, contre 44% et 19%, voir annexe D). L'« éboulis » du pourcentage d'inertie selon les dimensions de l'ACP permet, en application du critère du coude, nous conduit à retenir les 2 premières composantes principales (cf. annexe D).

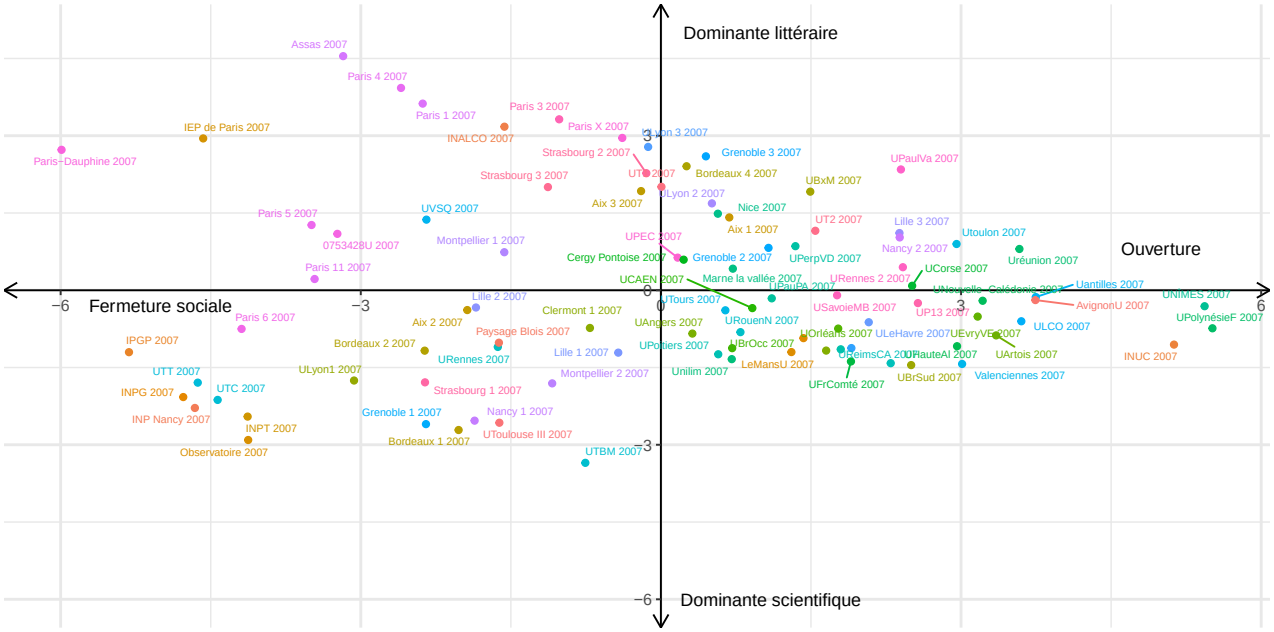
Fig. 1 – Représentation des variables de l'ACP sur les données de 2007



Note : Les variables actives sont représentées en noir et les variables supplémentaires en gris

Comme on peut le voir sur le graphique **??**, les deux dimensions principales de notre analyse correspondent bien à celles de l'ACP que nous cherchons à reproduire [Avouac et Harari-Kermadec 2022], avec un axe horizontal (première dimension) qui correspond à la composition sociale des universités, avec les publics d'origine favorisé à gauche et les publics d'origine populaire à droite, et un axe vertical (seconde dimension) qui oppose les établissements à dominante scientifique de ceux à dominante littéraire et/ou économique et sociale.

Fig. 2 – Représentation des individus de l'ACP en 2007



5.2 Projection des données de 2021 sur l'ACP des données de 2007

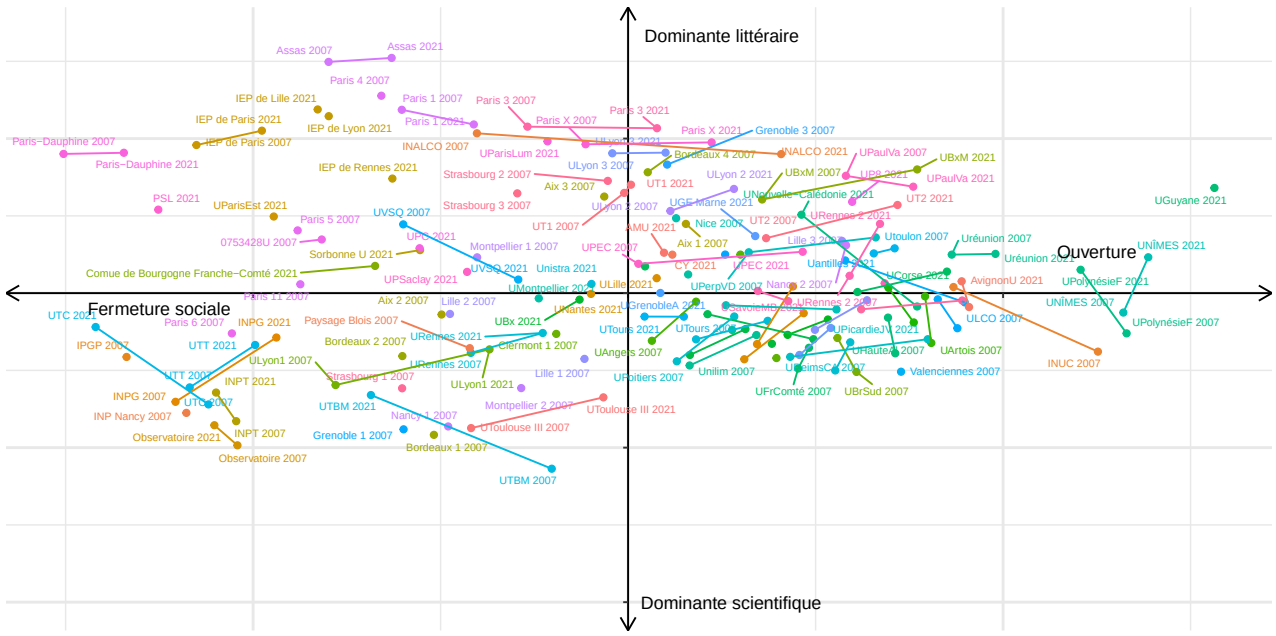
Il s'agit désormais de projeter sur notre ACP de 2007 les établissements de 2021, afin de pouvoir comparer en toute rigueur et sur le même plan factoriel l'évolution des positions des universités entre 2007 et 2011.

> Encadré 1 — Transformations réalisées sur la base `Info.Etab`
>
> Étant donné les nombreuses fusions d'établissements entre 2007 et 2021, il nous a fallu opérer quelques modifications sur la base de données `Info.Etab` de façon à y inclure les noms et sigles des établissements ayant été absorbés entre-temps. Il manquait de fait dans la base une quinzaine d'université (notamment Bordeaux 1, 2 et 3, Lille 1, 2 et 3, Strasbourg 1, 2 et 3, Paris 4, 5, 6, 7 et 11, Aix 1, 2 et 3 ou encore CergyPontoise) qui n'étaient donc pas représentées sur l'ACP.

Notre ACP, comme la précédente, est construite à partir de 13 variables, auxquelles s'ajoutent 5 variables inactives. Les variables et individus actifs sont les mêmes que pour l'ACP précédente, mais on a ajouté en individus supplémentaires les universités telles qu'elles étaient en 2021. Si le graphique des l'inertie en fonction des dimensions de l'ACP (cf. annexe D) est globalement très similaire à celui de l'ACP portant uniquement sur les données de 2007, on observe quelques différences, que nous ne parvenons pas à expliquer.

Le graphique 3 montre un certain nombre de changements au sein des universités françaises en 2007 et 2021, bien qu'il soit difficile d'identifier à ce stade des tendances globales, on peut distinguer un mouvement vers la droite (donc vers l'ouverture sociale) des universités du quart nord-ouest (public favorisé, dominante littéraire). Il semble plus difficile d'identifier une tendance pour les moitié droite (ouverture sociale, public populaires) et sud (dominante scientifique) du graphique.

Fig. 3 – Représentation des individus de l'ACP en 2007 et en 2021



Note : le plan est celui de l'ACP de 2007, les établissements de 2021 sont des individus supplémentaires.

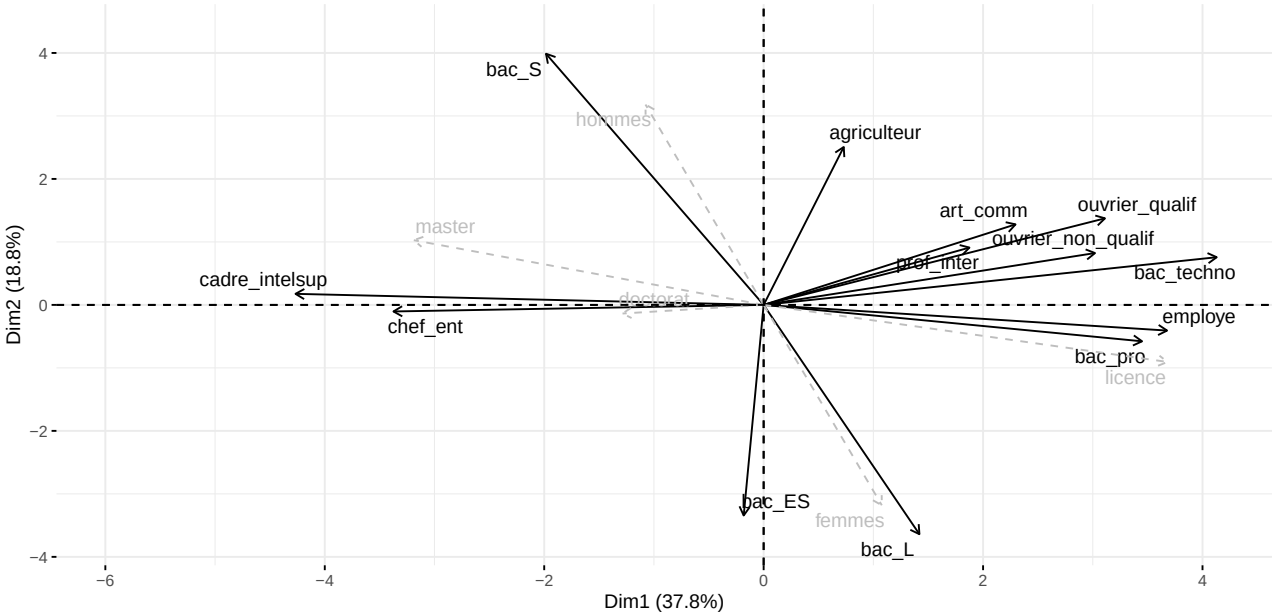
Deux hypothèses nous semblent à même d'expliquer ces changements dans la position des universités sur l'ACP entre 2007 et 2021 : il est possible que les axes de l'ACP de 2007 ne soient plus adéquats, que ces changements soient liés à des modifications structurelles, par exemple à la perte de pertinence de certaines variables. Si ce n'est pas le cas, alors la modification des positions sur le graphique pourrait résulter de changements dans la composition sociale des universités qui auraient occasionné entre 2007 et 2021 une recomposition dy paysage universitaire français. Nous allons réaliser une ACP portant uniquement sur les universités de 2021 afin de savoir si nos variables sont aussi pertinentes pour analyser les universités en 2021 qu'en 2007.

5.3 Réalisation d'une ACP sur les données de 2021

La graphique 4 montre que, selon toute vraisemblance, les variables utilisées sont aussi pertinentes en 2021 qu'en 2007. Elles suivent de surcroît une répartition analogue sur le graphique [2]. Ceci nous permet de conclure à des changements dans la composition sociale des universités, hypothèse qu'il s'agira d'approfondir et de préciser en menant une classification ascendante hiérarchique.

2. Le graphique présentant les positions des établissements sur l'ACP de 2021 (graphique 17) est disponible en annexe D

Fig. 4 – Représentation des variables de l'ACP sur les données de 2007



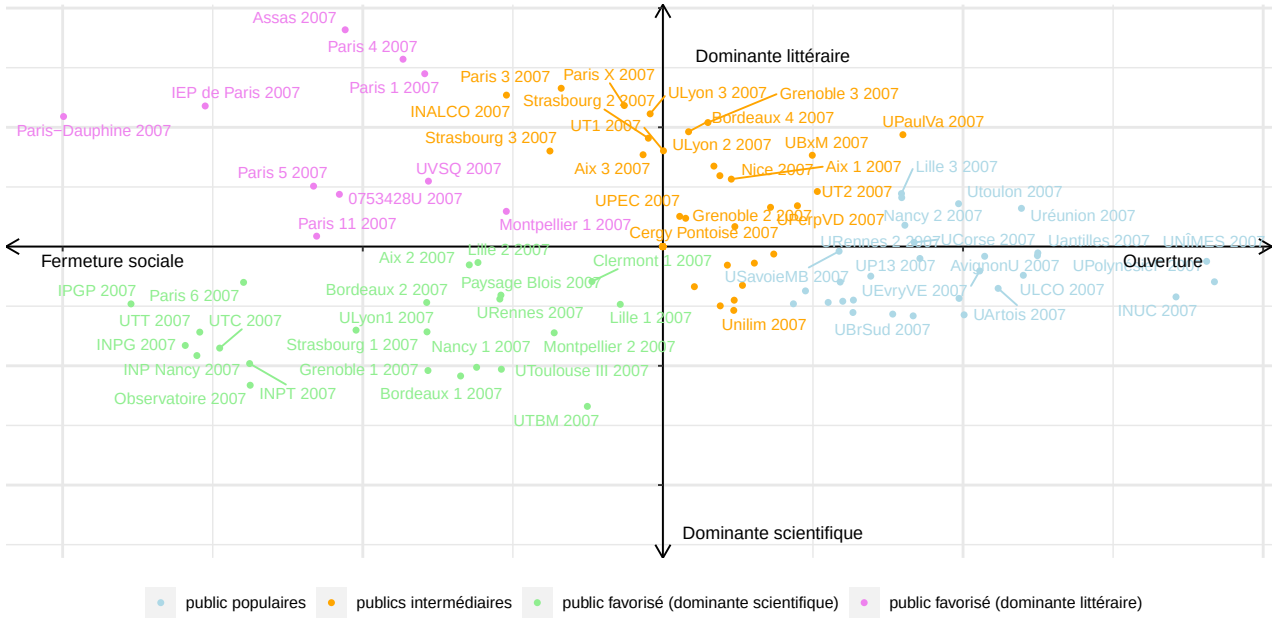
Note : les variables actives sont représentées en noir et les variables supplémentaires en gris.

6 Classification ascendante hiérarchique des universités

6.1 Essai de reproduction de la CAH présentée dans l'article

Il s'agit ici de reproduire la classification ascendante hiérarchique présentée dans l'article de Avouac et Harari-Kermadec 2022. La classification à laquelle nous parvenons (cf. graphique 5) est globalement similaire à celle que nous voulons reproduire, les différences s'expliquent principalement par un nombre beaucoup plus faible de points (d'universités observées). Nos groupes sont donc moins « consistants » que ceux de la CAH de Avouac et Harari-Kermadec, mais il nous semble que cela ne remet pas en cause la déclinaison des groupes qu'ils proposent, à savoir : public favorisés à dominante respectivement littéraire et scientifique, publics intermédiaires et publics populaires.

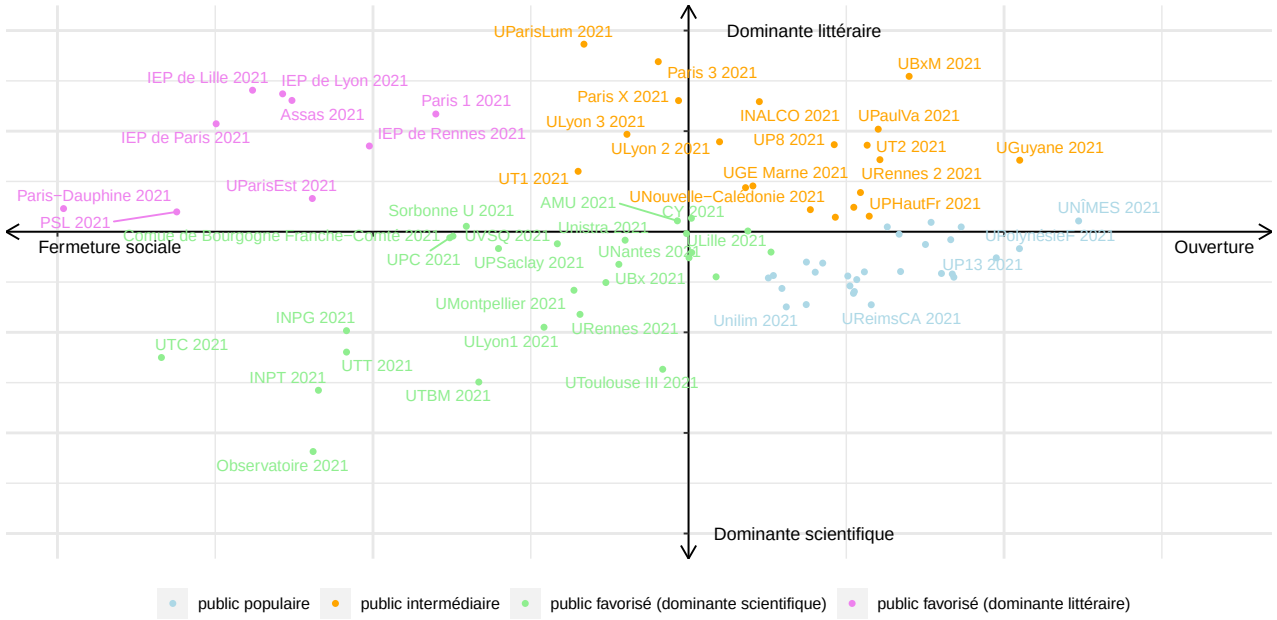
Fig. 5 – Représentation de la typologie issue de la CAH de 2007



6.2 CAH sur les données de 2021

Nous réalisons désormais une classification ascendante hiérarchique sur le modèle de la précédente pour les données établissements de 2021. Il s'agit de voir si nos groupes sont aussi « consistants » qu'en 2007, en sommes si une classification en quatre groupes est toujours pertinente en 2021 et quelles sont les modifications au sein des groupes.

Fig. 6 – Représentation de la typologie issue de la CAH de 2021



De fait, il semblerait au vu du graphique 6 que la déclinaison en quatre groupes adoptée pour 2007 demeure pertinente en 2021, avec deux groupes à composition sociale favorisée, qui regroupent des établissements à dominante respectivement scientifique et littéraire ou économique, un groupe à composition sociale intermédiaire et un dernier groupe accueillant un public populaire.

6.3 Dynamique des groupes d'établissements entre 2007 et 2021

Afin de comparer l'évolution de chaque groupe sur le plan entre 2007 et 2021, nous représentons l'évolution du barycentre de chaque groupe sur le plan factoriel entre 2007 et 2021 (les points moyens de chaque groupe sur les deux dimensions de lACP, pondérés par le nombre détudiants inscrits dans les différents établissements). Pour plus de détails, voir l'encadré 2.

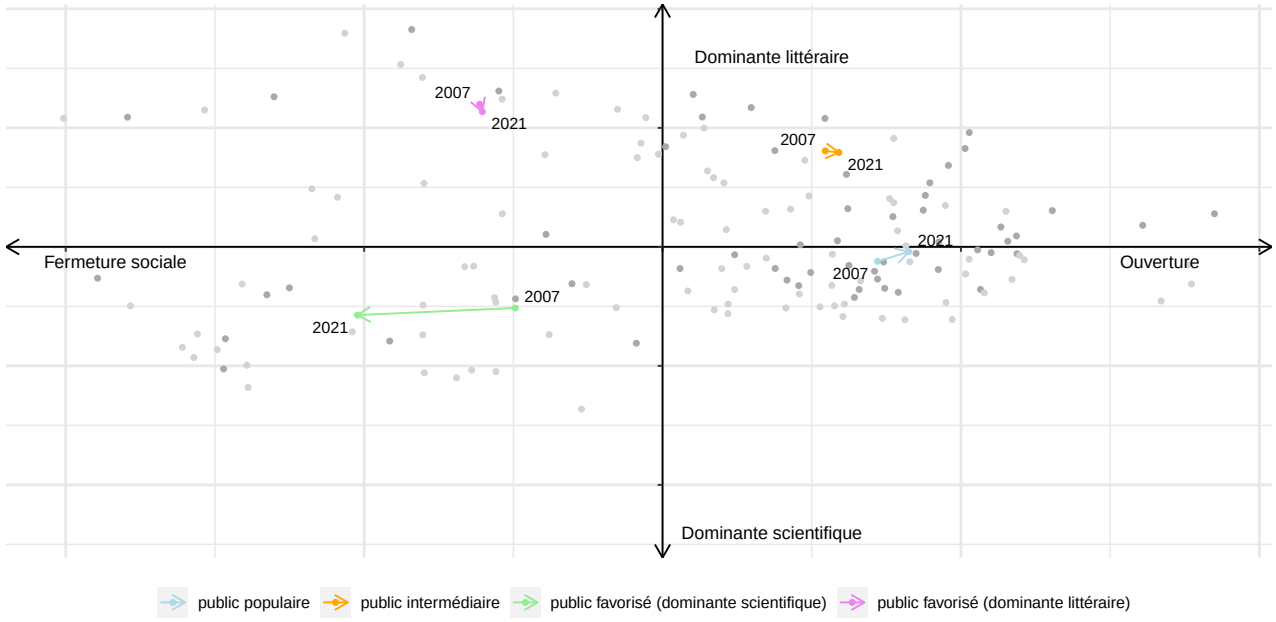
Le groupe des établissements à dominante scientifique et à composition sociale favorisée paraît se fermer fortement, pour accueillir un public encore plus favorisé en 2021 qu'en 2007 tandis que les groupes accueillant des publics populaires et intermédiaires semblent eux aller vers une ouverture sociale encore plus grande (cf. graphique 7). Ces quelques observations semblent attester d'une dynamique des extrêmes, avec un écart croissant entre les établissements favorisés et les établissements populaires. Rappelons néanmoins que sur ces dynamiques ne prennent pas en compte certains établissements importants tels que PSL ou Paris-Saclay, créés entre 2007 et 2021.

Encadré 2 — Méthode pour mesurer l'évolution de la position de chaque groupe sur le plan de l'ACP

Comme pour la comparaison intertemporelle des positions des établissements sur l'ACP, nous reprenons le plan de l'ACP de 2007 sur lequel nous projetons les établissements tels qu'ils étaient en 2021. La difficulté est ici de savoir à quel groupe appartiennent les établissements de 2021. Nous utilisons la variable `Etablissement` afin, dans le but de comparer des groupes qui soient les plus homogènes possibles entre 2007 et 2021, de placer les établissements de 2021 dans e groupe qui était le leur en 2007. Cette méthode a des inconvénients, notamment parce qu'elle exclut les établissements nés de fusions qui n'apparaissent donc pas dans la base de 2007, ce qui est d'autant plus regrettable que ces établissements sont souvent d'un taille conséquente (c'est le cas pour l'université PSL ou Paris-saclay Université, pour ne citer qu'elles). Soulignons néanmoins que ces inconvénients sont inhérents à toute comparaison intertemporelle : même en procédant à un codage manuel, il serait impossible de faire figurer ces nouveaux établissements dans un groupe donné sans opérer de choix normatifs, étant donné que les différents établissements composantes de ces grandes universités n'appartenaient le plus souvent, en 2007, pas à seul et unique groupe.

Par ailleurs, nous appuyer sur les groupes issus de la CAH de 2021 n'aurait pas davantage de sens étant donné que nous comparerions alors des points situés sur des plans différents (le plan de l'ACP de 2007 et celui de l'ACP de 2021) et, de surcroît, les barycentres de groupes différents, certains établissements étant passés d'un groupe à l'autre en 14 ans.

Fig. 7 – Évolution des barycentres dans la typologie à quatre groupes issue de la CAH entre 2007 et 2021



Note : les établissements de 2007 apparaissent en gris clair sur le graphique, ceux de 2021 en gris foncé.

Références

[1] Romain Avouac. *Mémoire d'initiation à la recherche On the performative e ects of university rankings : evidence of increasing social polarization among French universities*. 2017.

[2] Romain Avouac et Hugo Harari-Kermadec. « French Universities A Melting Pot or a Hotbed of Social Segregation ? A Measure of Polarisation within the French University System (2007-2015) ». In : *Economie et Statistique / Economics and Statistics* 528-529 (5 jan. 2022), p. 63-83.

[4] Joan-Maria Esteban et Debraj Ray. « On the Measurement of Polarization ». In : *Econometrica* 62.4 (juill. 1994), p. 819.

[9] Thierry Rocher. « Construction d'un indice de position sociale des élèves ». French. In : *Éducation & formations* 90 (avr. 2016), p. 5-27.

Logiciels et paquets

[3] David B. Dahl, David Scott, Charles Roosen, Arni Magnusson et Jonathan Swinton. *xtable : Export tables to LaTeX or HTML*. manual. 2019.

[5] Julien Gossa. *kpiESR : DataViz for DataESR*. manual. 2023.

[6] Marek Hlavac. « stargazer : Well-Formatted Regression and Summary Statistics Tables ». In : (2018).

[7] Christian Kleiber et Achim Zeileis. *Applied econometrics with R*. New York : Springer-Verlag, 2008.

[8] R Core Team. *R : A language and environment for statistical computing*. manual. Vienna, Austria : R Foundation for Statistical Computing, 2022.

[10] Hadley Wickham. *ggplot2 : Elegant graphics for data analysis*. Springer-Verlag New York, 2016.

[11] Hadley Wickham, Jim Hester et Jennifer Bryan. *readr : Read rectangular text data*. manual. 2023.

[12] Hadley Wickham et al. « Welcome to the Tidyverse ». In : *Journal of Open Source Software* 4.43 (21 nov. 2019), p. 1686.

[13] Yihui Xie. « knitr : A comprehensive tool for reproducible research in R ». In : *Implementing reproducible computational research*. Sous la dir. de Victoria Stodden, Friedrich Leisch et Roger D. Peng. Chapman and Hall/CRC, 2014.

[14] Hao Zhu. *kableExtra : Construct complex table with 'kable' and pipe syntax*. manual. 2021.

Annexes

Annexe A - Informations sur les variables

Tableau 3 – Informations générales sur les principales variables

	min.	Q1	médiane	moyenne	Q3	max.	NA
effectif	68.00	7507.00	16962.00	21031.04	30266.00	73598.00	1
Init_Excel	0.00	0.00	0.00	0.29	0.00	2.00	0
public_défavorisé	0.08	0.19	0.22	0.23	0.27	0.53	1
master	0.00	0.20	0.27	0.33	0.36	1.00	1
doctorat	0.00	0.02	0.02	0.05	0.04	0.91	1
bac_S	0.13	0.30	0.40	0.40	0.47	1.00	1
femmes	0.18	0.54	0.58	0.57	0.62	0.77	1
class_shanghai	13.00	800.00	1100.00	886.42	1100.00	1100.00	13
cohorte_shanghai	0.00	5.00	6.00	5.30	6.00	6.00	0

Annexe B - Suppléments concernant la régression linéaire

Fig. 8 – Régression linéaire avec l'effectif comme seule variable explicative

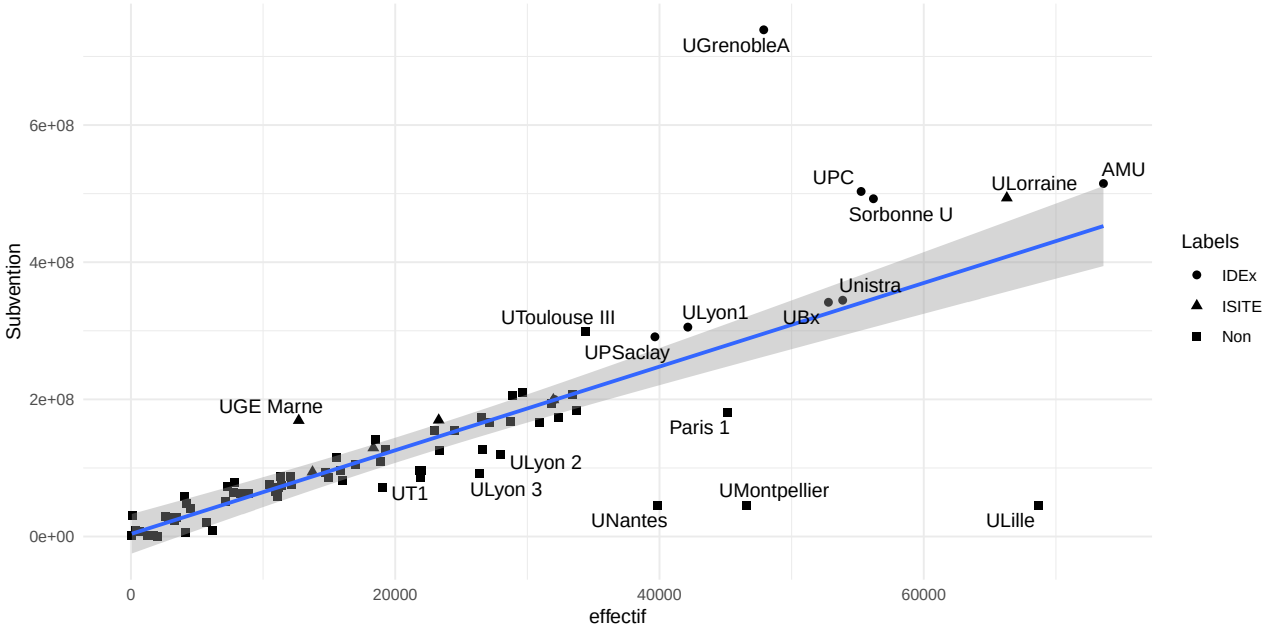
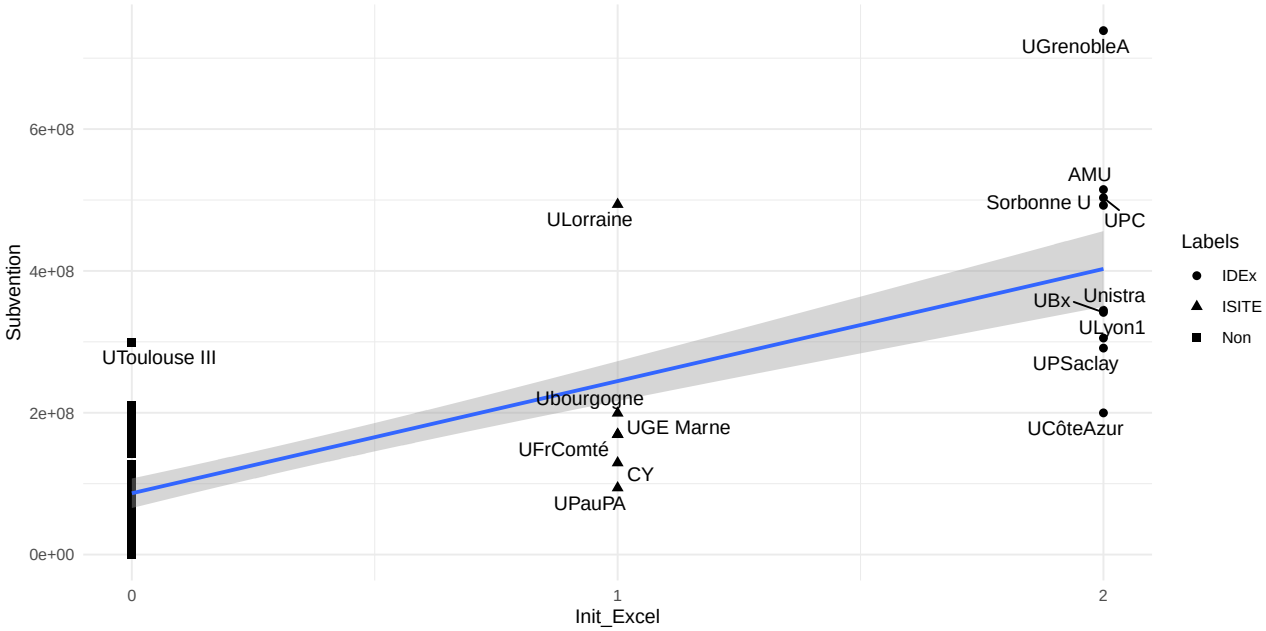


Fig. 9 – Régression linéaire avec `Init_Excel` comme seule variable explicative



Annexe C - Suppléments concernant le modèle Tobit

FIG. 10 – Modèle tobit avec comme unique régresseur la proportion d'étudiant·es défavorisé·es

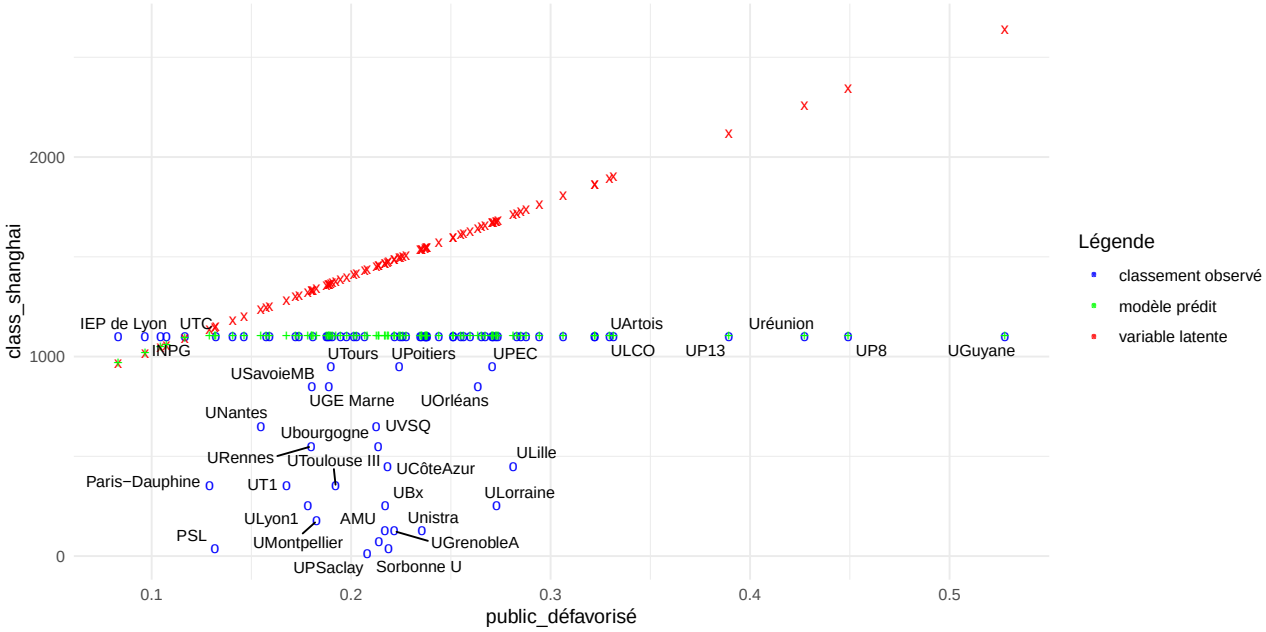


FIG. 11 – Modèle tobit avec comme unique régresseur la proportion de bac S

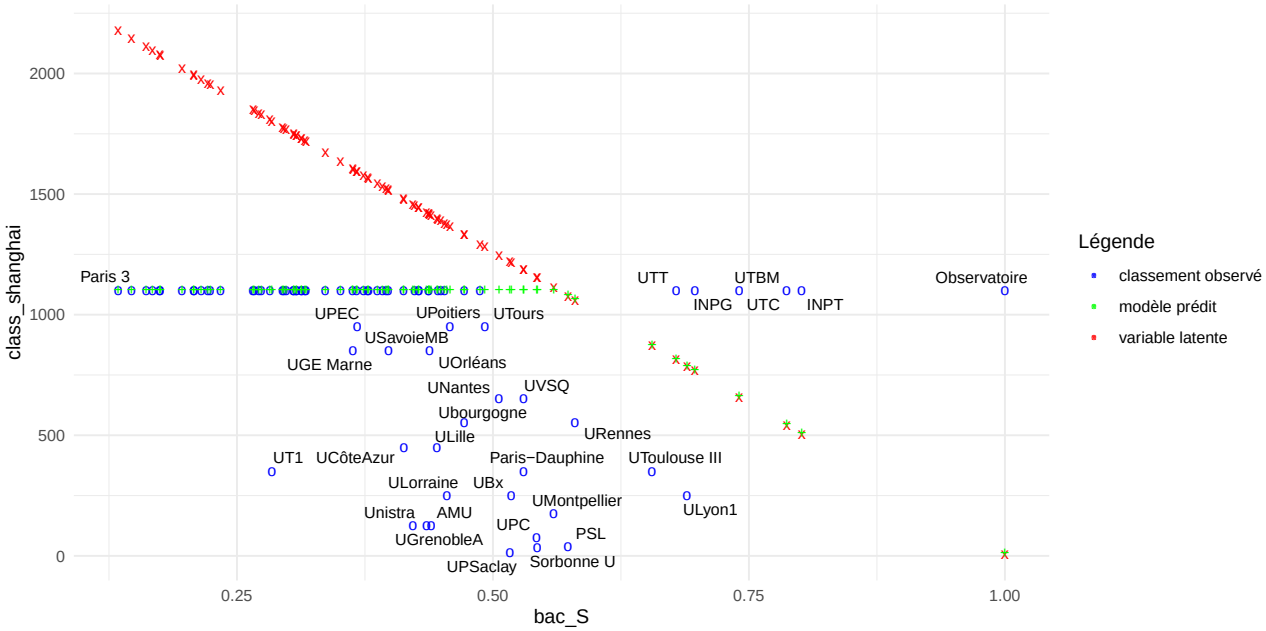


Fig. 12 – Modèle tobit avec comme unique régresseur l'effectif des établissements

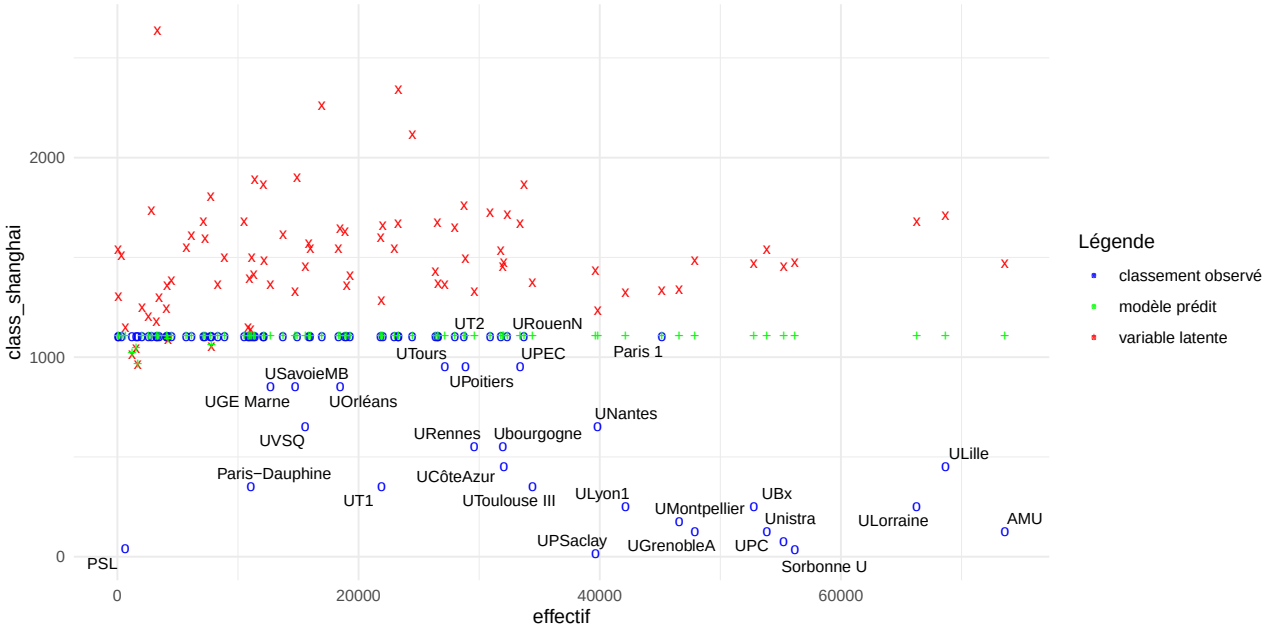
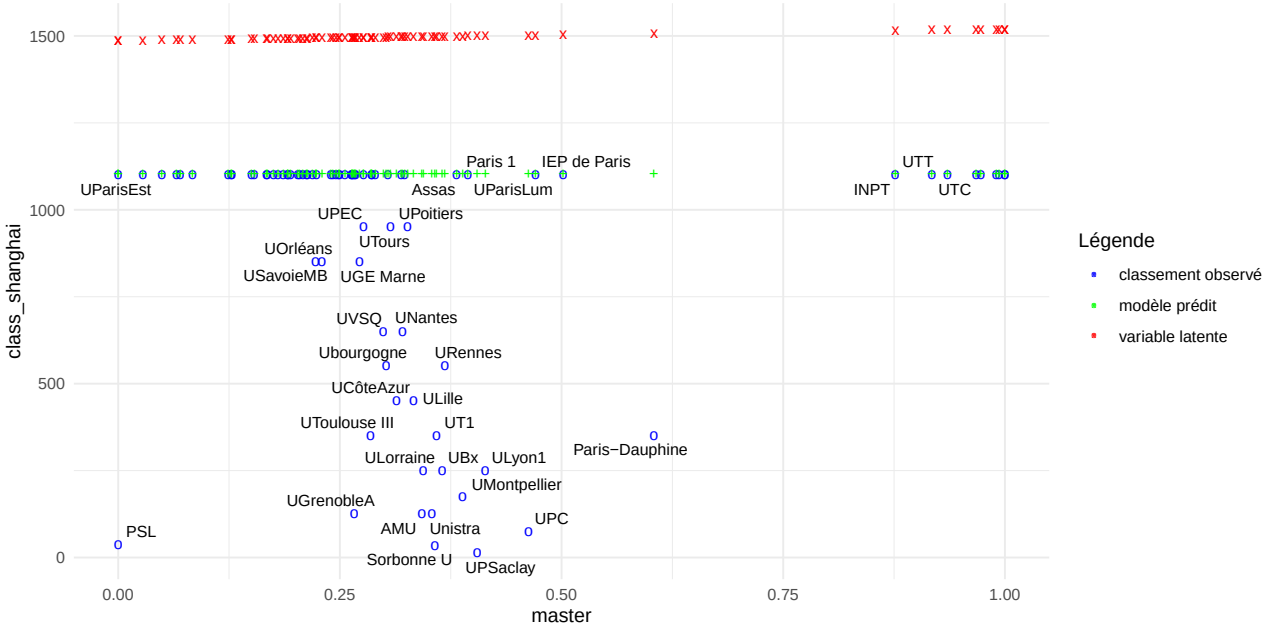


Fig. 13 – Modèle tobit avec comme unique régresseur le taux d'étudiant·es en master



Annexe D - Suppléments concernant l'ACP

FIG. 14 – Variance expliquée par les différentes dimensions de l'ACP en 2007

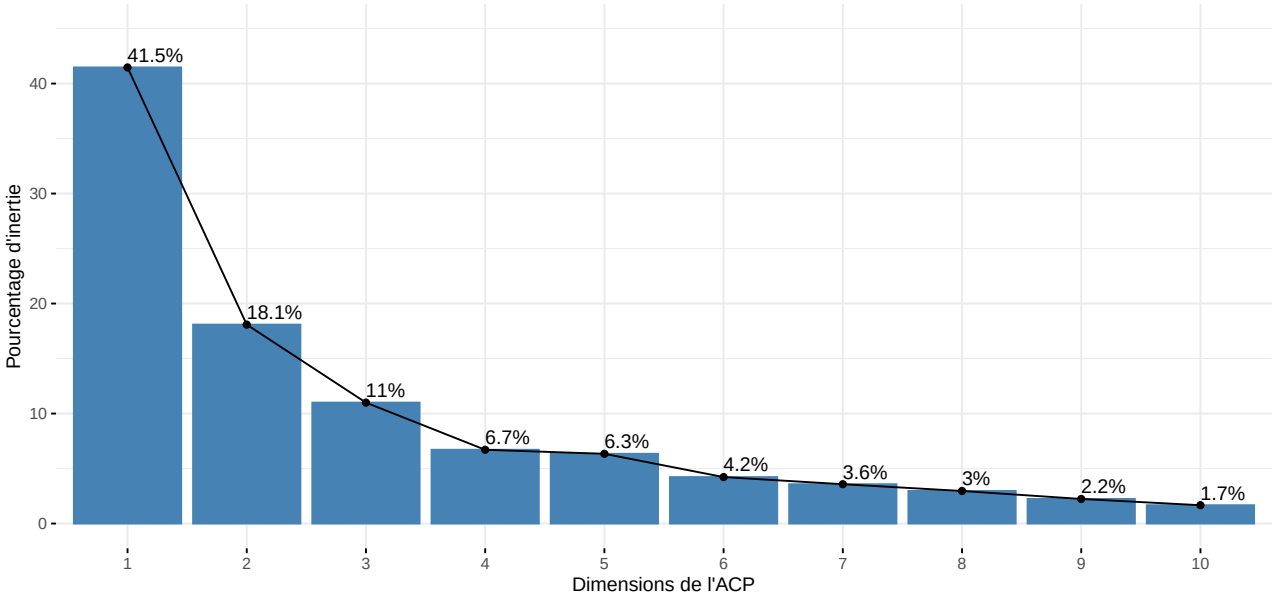


FIG. 15 – Variance expliquée par les différentes dimensions de l'ACP en 2007, avec en individus supplémentaires les établissements de 2021

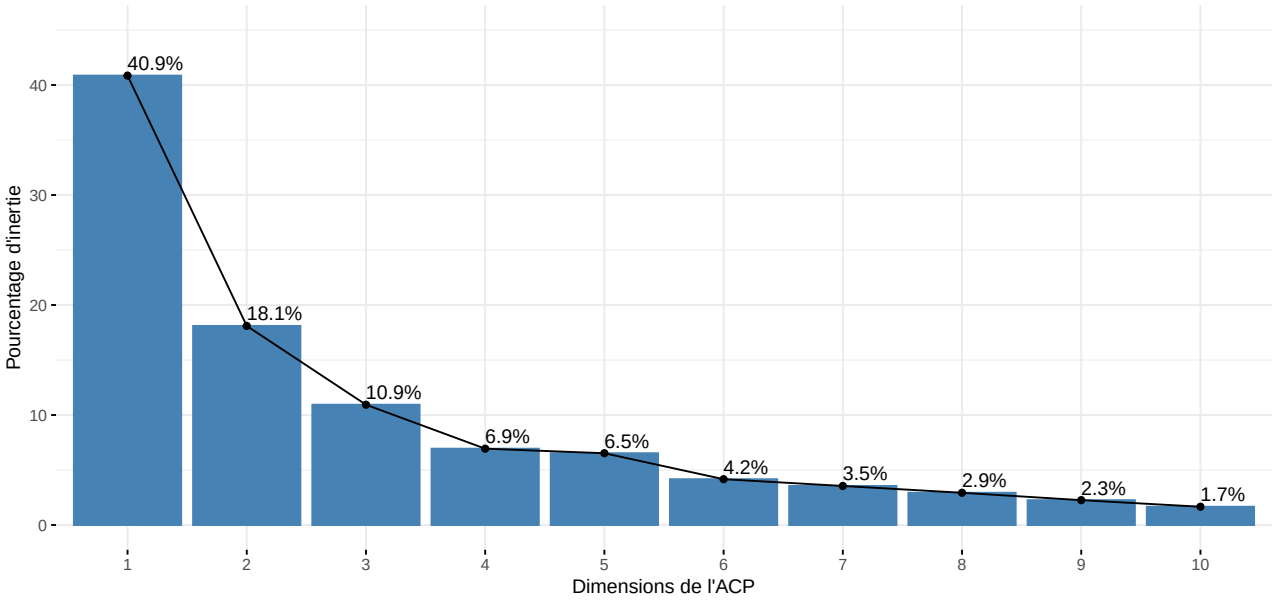


FIG. 16 – Projection des individus de 2021 sur le plan de l'ACP de 2007

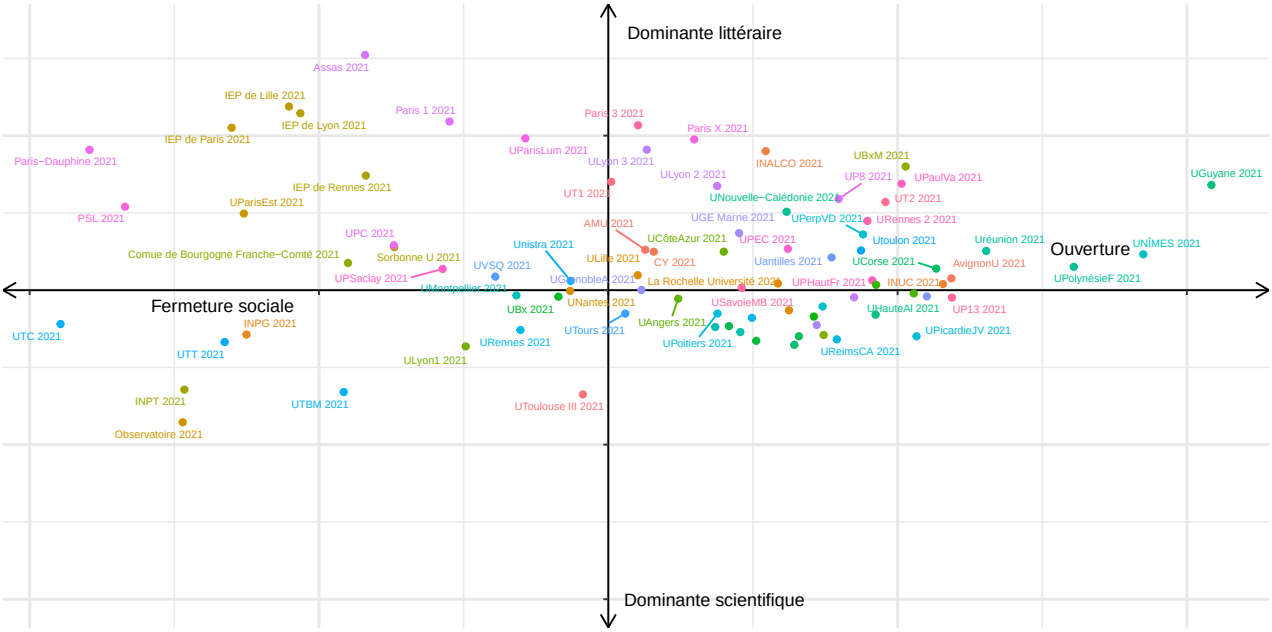
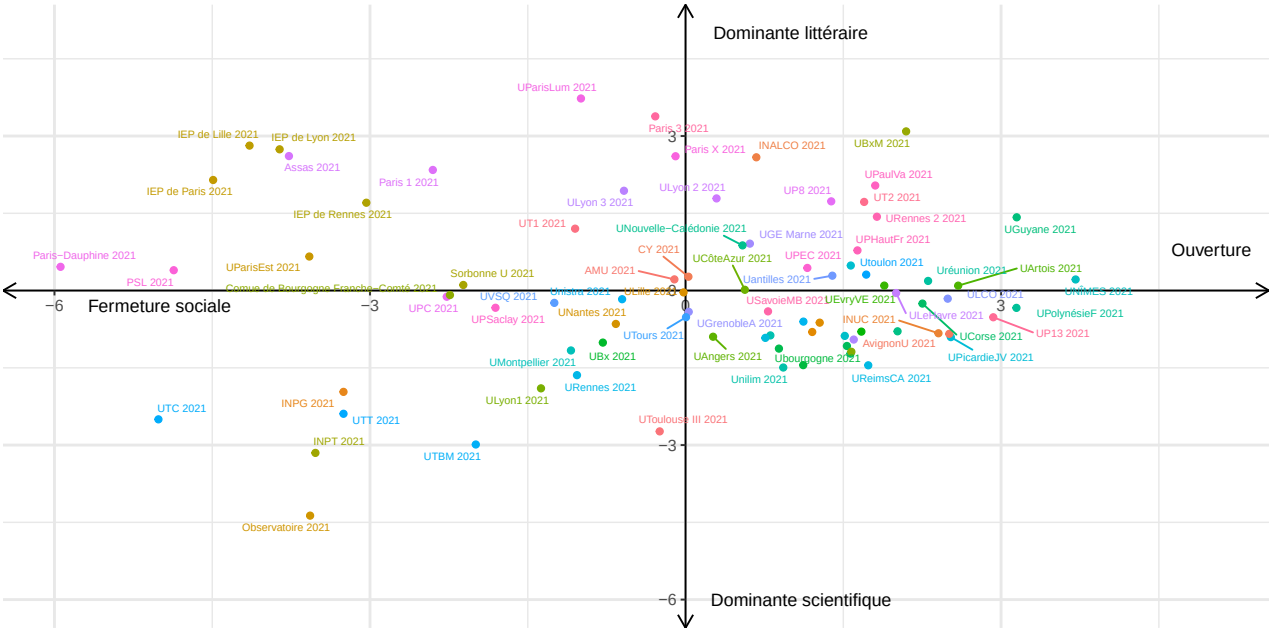
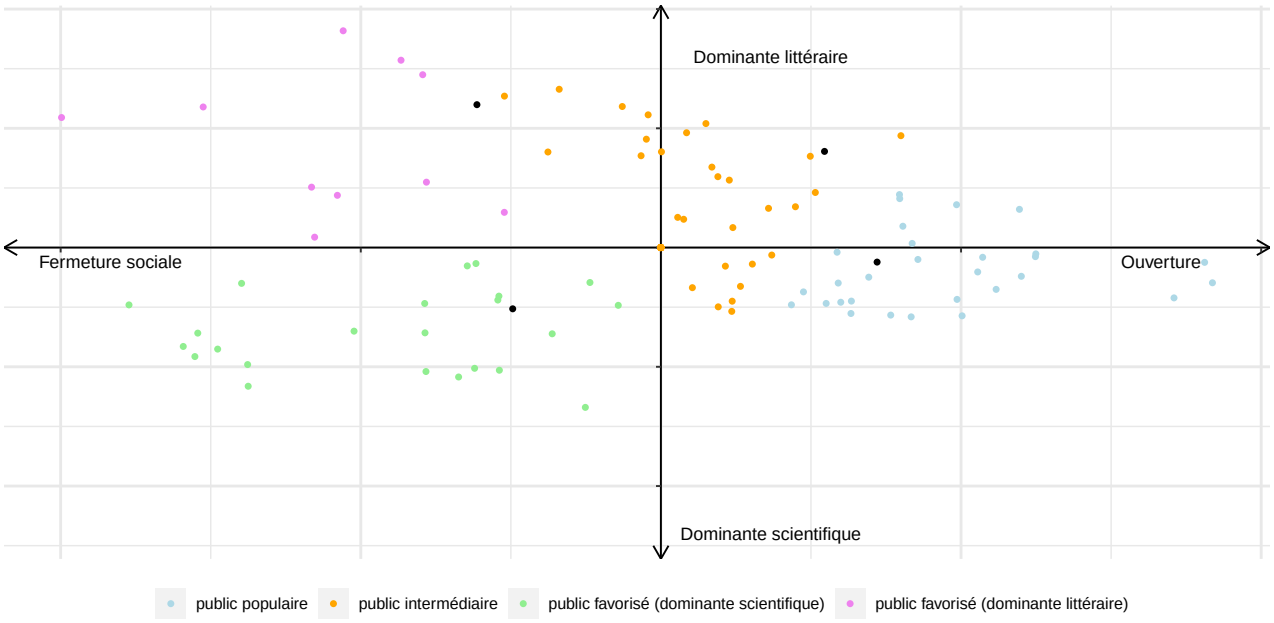


FIG. 17 – Représentation des individus de l'ACP en 2021



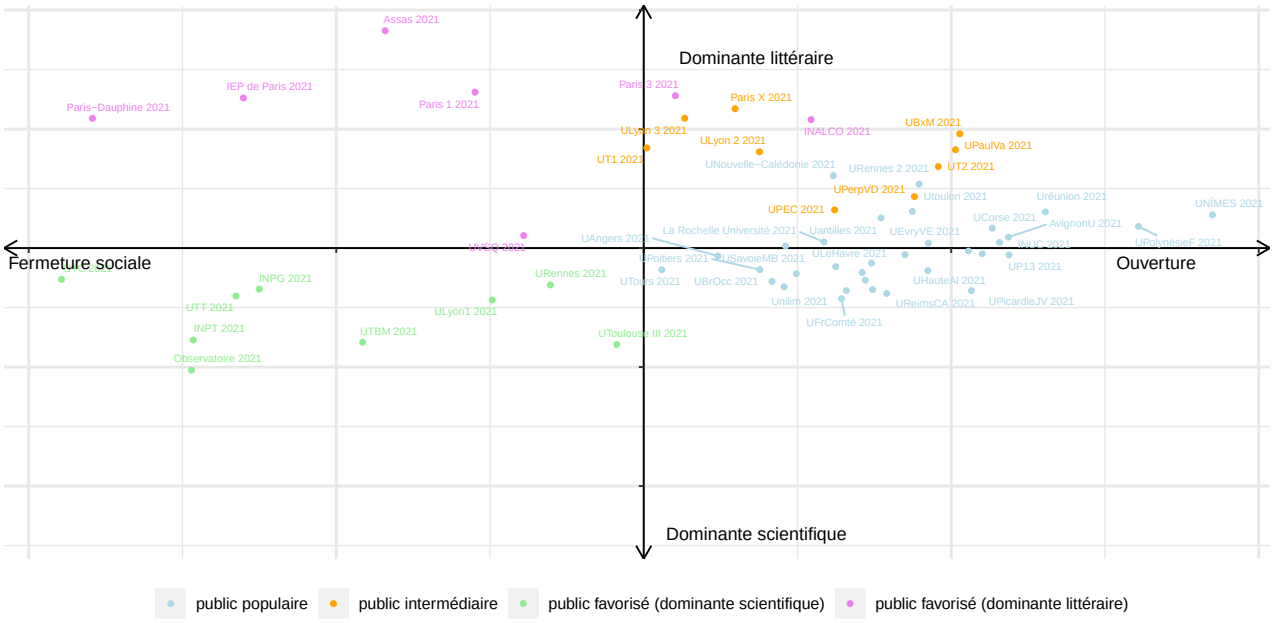
Annexe E - Suppléments concernant la CAH

Fig. 18 – Représentation des barycentres de la typologie à quatre groupes issue de la CAH de 2007



Note : en noir apparaît le barycentre de chacun des groupe.

Fig. 19 – Projection des individus de 2021 sur le plan de l'ACP de 2007 en conservant la typologie de la CAH de 2007



Note : le groupe d'appartenance de chaque établissement est celui issu de la CAH de 2007, les établissements créés depuis 2007 ont été ôtés du graphique.