Dans quelles mesures l'origine sociale est un déterminant du choix des langues au collège ?

Théophile FROMENT et Julia NICOLAS

Abstract

French schooling is often the focus of attention when it comes to issues of inequality. The middle school in particular is often criticized and is subject to reforms that attempt to reduce inequalities and what is called school segregation. It is precisely in middle school that the choice of the modern languages that students will study throughout their schooling takes place. It is noted that certain languages are socially marked. Many sociological field analyses or descriptive surveys have shown that the choice of certain languages can appear as a class strategy. Few econometric studies have been carried out in this sense and the objective of this work is to determine quantitatively whether a student's CSP has an impact on his/her choice. We focused on the choice distinction between German and Spanish. We show that positive effects are noted between the choice of german over spanish when more higher csp chose german. However, this study was not sufficient to explain all the parameters that make up student choice.

Remerciements : Nous tenons à remercier en premier lieu Monsieur Roland RATHELOT pour l'encadrement de ce projet. Ses nombreux conseils et recommandations nous ont permis de développer notre réflexion sur la réalisation d'un projet de recherche de la question de recherche à la stratégie empirique. Nous voulons également remercier les deux autres groupes avec lesquels nous avons échangé idées et présentations tout au long de l'année.

Introduction

Dès leur entrée à l'école primaire, les enfants d'origines sociales défavorisées sont souvent victimes d'inégalités en terme de résultats scolaires. En effet, les résultats obtenus par ces élèves sont en moyenne inférieurs à ceux des enfants de cadre, et ces inégalités perdurent au collège : les enfants d'ouvriers, d'employés et d'inactifs représentent 86 % des élèves des sections d'enseignement général et professionnel adapté, contre 53 % dans l'enseignement général [3]. Souvent perçu comme un tournant important dans la scolarité des élèves, le collège est régulièrement critiqué pour ses facteurs qui désavantageraient les élèves de certaines catégories sociales et creuseraient encore plus les écarts de niveau. Dans ce projet, nous nous intéressons plus particulièrement à un des premiers choix que les élèves doivent faire durant leur scolarité : celui d'une langue ou deux à étudier à partir de la 6e et de la 4e qu'ils garderont jusqu'au baccalauréat en Terminale. En France, les langues proposées à l'étude sont en majorité l'anglais, l'espagnol et l'allemand et le choix entre toutes découle souvent de plusieurs facteurs difficilement mesurables, comme l'affinité pour une culture ou l'utilité perçue de l'apprentissage d'une certaine langue. Cependant, il semble déjà exister une certaine

différenciation sociale dans les choix des élèves entre ces langues [1]. L'objectif de ce travail sera de se demander dans quelles mesures l'origine sociale peut être un déterminant du choix des langues vivantes au collège. Nous avons choisi de travailler sur la distinction LV2 allemand et LV2 espagnol. Pour cela nous avons créé un modèle binaire de choix influencé par la population antérieure dans les différentes classes de langue au sein d'un même collège. Cette étude a été réalisé grâce à des données portant sur l'ensemble des élèves dans la région île-de-France entre 2006 et 2018.

1 Contexte et justification

1.1 Le constat de langues socialement marquées

L'idée de notre étude vient d'un constat d'une étude du CNESCO (Centre National des Études du Système Scolaire) de 2019 sur les choix de langues au collège en France [1]. Cette étude révèle que les choix de langues étrangères sont socialement marqués à la fois en LV1 comme en LV2. Par exemple, les élèves les plus favorisés sont surreprésentés en LV2 allemand, anglais ou chinois. Plus généralement, dans la sociologie de l'éducation, des études démontrent que l'offre scolaire (donc l'offre de langues) participe à des phénomènes de ségrégation sociale. En effet, d'après [4], il existe une forte corrélation entre le profil socio-économique et socioculturel des communes et les offres scolaires de leurs collèges. Par exemple, les collèges des communes plus favorisées bénéficient d'une plus grande richesse en options (sections européennes et internationales, etc) par rapport à ceux des communes populaires. Ce constat est partagé par François Baluteau [2], qui montre que les collèges s'ordonnent également selon la composition sociale. Si plus de la moitié des collèges les plus «favorisés» propose trois langues, plus d'un tiers en offre quatre et un dixième au moins cinq. En revanche, les collèges «défavorisés» réduisent leur offre à trois langues vivantes pour les deux tiers et à quatre pour un quart d'entre eux. Les langues vivantes seraient donc caractéristiques d'un certain indicateur de standing qui bénéficierait aux classes sociales supérieures.

1.2 L'hypothèse d'une stratégie d'évitement ou de regroupement

Beaucoup d'hypothèses peuvent expliquer ces disparités sociales au sein des choix des langues, mais aucune ne semble avoir été prouvée empiriquement. On pourrait par exemple supposer que des langues en particulier sont perçues comme étant plus exigeantes que d'autres, et qu'alors certains parents de classes sociales favorisées encouragent leurs enfants à étudier ces langues. On sait également que dans de nombreux collèges, les contraintes au niveau des effectifs et des horaires obligent les établissements à créer des classes qui regroupent les élèves étudiant la même langue : ils sont alors réunis pour l'ensemble des autres cours. Dans ce cas, les choix des langues pourraient apparaître comme stratégiques pour certains parents favorisés, qui pensent que le choix d'une langue conduit leur enfant à étudier avec des élèves du même milieu social. C'est d'ailleurs la crainte d'une stratégie de regroupement des classes sociales élevées dans les mêmes classes qui avait initié l'adoption d'une réforme des collèges et des langues vivantes en 2015, limitant ainsi l'existence de classes bilangues dès la 6e. Pour caractériser l'influence de la catégorie sociale sur le choix de certaines langues, nous pouvons alors nous demander si la population dans les différentes classes de langue peuvent permettre aux parents de les influencer dans le choix de la LV2 pour leurs enfants.

2 Les données

Nous cherchons donc dans ce projet à apporter des éléments de réponses quantitatifs et économétriques à tous ces travaux. Nous avons alors fait les demandes pour récupérer les données de la Base Centrale Scolarité (BCS) de 2005 à 2018. Ces données couvrent 99% des élèves des établissements du

second degré publics et privés (sous contrat et hors contrat) dépendant du Ministère de l'éducation nationale (France métropolitaine + DOM). Deux tables par année sont données :

- une table **élèves** : descriptions très détaillée de chacun des élèves (variables sociodémographiques comme l'âge, le sexe, l'origine sociale des parents, la nationalité (à l'échelle du continent), mais également la classe, le parcours scolaire, etc).
- une table **établissements** : chacun est distingué par un numéro d'établissement (également présent dans la base élèves, pour pouvoir identifier le lieu d'études), avec des informations sur privé / public, type d'établissement, zone géographique, etc.

3 Stratégie empirique

3.1 Statistiques descriptives et travaux préliminaires sur la base

À partir de nos données, nous voulons dans un premier temps retrouver les résultats de l'étude du CNESCO. À l'échelle descriptive, quelle est la répartition des classes socio-professionnelles dans les choix des élèves en terme de langues ? En France, la plupart choisissent d'étudier l'anglais en tant que LV1 (Figure 1). Nous nous pencherons donc plutôt sur le choix de la LV2 qui sont au nombre de 16. Pour la LV2, les différents choix sont répartis de manière plus uniforme, même si l'espagnol reste la langue majoritairement choisie dans notre échantillon (Figure 2). À partir des résultats de nos statistiques descriptives (4.1), nous relèverons plusieurs langues marquées socialement, c'est-à-dire où le taux de CSP+ est plus ou moins élevé que la moyenne globale.

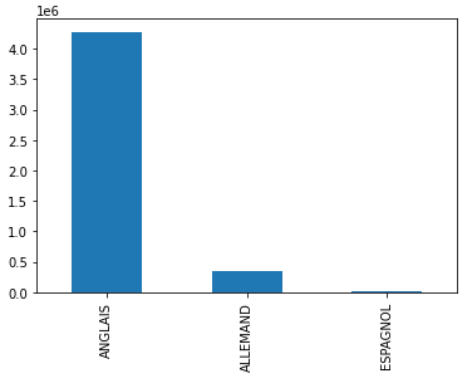

Figure 1: Répartition du nombre d'élèves dans chaque LV1

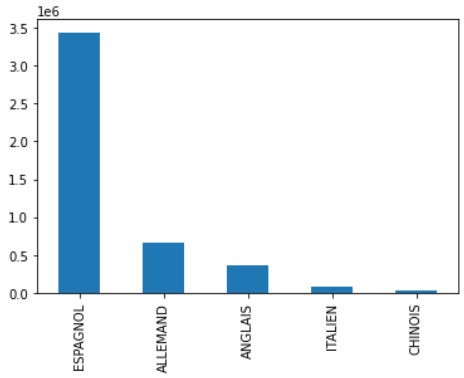

Figure 2: Répartition du nombre d'élèves dans chaque LV2

Nous avons ensuite cherché un cadre précis à notre étude sur les déterminants des choix de la langue. Nous avons réduit la base d'étude aux collèges en Île de France. En effet, c'est au collège que le choix de LV2 a lieu, et l'élève est généralement contraint de maintenir ce choix à son entrée au lycée. Au vue de la taille de la base, il était aussi plus intéressant de réduire l'étude à une certaine zone géographique. Nous avons choisi l'Île de France pour plusieurs raisons. Premièrement, région très peuplée, c'est elle qui offre une répartition de collèges la plus concentrée, avec de nombreux collèges publics et privés de tailles et de populations différentes. Deuxièmement, nous nous intéressons au choix d'une certaine langue à apprendre. C'est un choix qui peut être justifié par le goût personnel pour une certaine culture (surtout s'il s'agit du choix d'une langue "secondaire", qui vient après le choix courant de l'anglais, souvent lié à des considérations d'utilité future dans le monde professionnel), ou encore par la proximité avec une zone géographique marquée par cette

culture linguistique. Dans ce sens, l'Île de France nous paraissait la meilleure option pour limiter cet effet non mesurable de l'appétence purement culturelle pour une langue (contrairement à des zones géographiques comme l'Alsace ou le Sud Ouest où l'on pourrait supposer que les élèves sont plus enclins à choisir respectivement l'allemand et l'espagnol en LV2).
Enfin, pour distinguer les différentes classes socioprofessionnelles et mettre en évidence des résultats sur les classes supérieures, nous avons utilisé les variables sur les métiers des responsables des différents élèves. Nous avons considéré qu'un élève appartenait à une classe sociale supérieure si au moins un de ses responsable légal avait un métier codé dans une catégorie CSP+ : chefs d'entreprises, artisans et commerçants, cadres, professions intellectuelles supérieures et professions intermédiaires.

3.2 Modèle binaire de choix entre LV2 espagnol et LV2 allemand

La deuxième étape de notre travail consiste à modéliser le choix de la LV2 pour un élève en classe de 4e à travers plusieurs variables explicatives en lien avec les catégories socioprofessionnelle des parents.
Pour ce faire, nous avons fait le choix d'une situation binaire où l'élève, à son entrée en quatrième, a le choix entre deux LV2 : l'espagnol et l'allemand. En effet, cette configuration est la plus courante dans l'ensemble de nos données. De plus, d'après les statistiques descriptives décrites plus bas (4.1), il apparaît que l'espagnol est plutôt une langue neutre en terme de marquage social alors qu'au contraire, l'allemand est quant à elle une langue où les classes supérieures semblent être sur représentées en général par rapport à la moyenne.

3.2.1 Les variables du modèle

Notre but est de mesurer si les catégories socioprofessionnelles ont un rôle dans le choix de la LV2, et comment elles peuvent agir.
Nous avons décidé de prendre comme variable explicative du choix de la LV2 un indice dit de **"dissimilarité"** caractérisant l'allemand dans un collège. L'indice de dissimilarité d'une langue est défini comme la différence δ_j entre la proportion de CSP+ dans la langue j du collège l'année précédant le choix, et la proportion globale de CSP+ dans le même collège toujours l'année précédant le choix. Autrement dit, dans notre cas l'indice mesure la part de la population de CSP+ qui devrait se déplacer pour que les deux langues soient représentatives de la population globale du collège. En choisissant cette variable, nous faisons l'hypothèse qu'avant de faire le choix, les parents d'élèves encouragent leur enfant à choisir la langue en fonction du nombre de CSP+ qui étudiaient la langue l'année précédente. Nous pouvons supposer un effet positif de la dissimilarité sur la probabilité de choisir allemand, effet qui pourrait évoluer aussi selon la CSP des parents de l'élève.
Nous disposons également de la variable binaire **CSP+** (au moins un des responsables appartient à une classe socioprofessionnelle supérieure). Nous nous attendons à un effet positif de la variable sur la probabilité de choisir allemand.
De plus, nous contrôlerons ce choix par la variable **Niveau** du collège l'année précédente. Le niveau d'un collège est défini comme la moyenne de son taux de réussite au brevet et de son taux d'obtention de mentions sur l'année précédente. Cette variable a pour objectif de contrôler l'effet qui pourrait subsister dans les collèges d'excellence académique où le choix de la langue n'influerait ni sur le niveau final au brevet ni sur la propension à côtoyer des CSP+.
Enfin, afin de mesurer autrement l'effet de la CSP des parents sur la probabilité de choisir LV2 allemand, nous avons introduit deux termes d'interactions entre les différentes variables suivantes:

- **CSP+** $*$ **niveau** permet de mesurer si l'effet d'appartenir à une CSP+ sur la probabilité de faire allemand en LV2 varie avec le niveau du collège. Si le coefficient devant cette variable

est négatif, cela signifie que plus le collège a un bon niveau, moins le fait d'avoir au moins un de ses parents CSP+ augmente la probabilité de choisir LV2 allemand. Cette dynamique peut sembler plutôt prévisible : on peut s'imaginer que dans les collèges les moins bons, les parents CSP+ ont tendance à préférer davantage que leurs enfants choisissent allemand LV2, langue réputée plus compliquée et "élitiste" que l'espagnol.

- **CSP+ $*$ dissimilarité** permet de mesurer si l'effet d'appartenir à une CSP+ sur la probabilité de faire allemand en LV2 varie avec l'indice de dissimilarité au sein du collège l'année précédant le choix. Si le coefficient devant cette variable est positif, cela signifierait que plus les CSP+ sont surreprésentés dans la LV2 allemand par rapport au taux global de CSP+ dans le collège, plus le fait d'avoir au moins un de ses parents CSP+ augmente la probabilité de choisir LV2 allemand. De même que pour le terme précédent, cette dynamique peut être intuitive : dans les collèges où les inégalités sont plus marqués (le taux de CSP+ dans les classes allemand est supérieur au taux global de CSP+ dans le collège), les parents CSP+ désirent peut-être davantage que leurs enfants choisissent LV2 allemand.

3.2.2 Le modèle Logit espagnol / allemand

Modèle binaire : $Y_i = \begin{cases} 1 & \text{si} LV2_i = \text{allemand} \\ 0 & \text{si} LV2_i = \text{espagnol} \end{cases}$ **Variable latente** $Y_i^* = X_i^* \beta + \epsilon_i$

3.2.3 Échantillon choisi

À travers notre indice de dissimilarité, nous cherchons dans ce modèle à évaluer si des variations de catégories sociales dans les différentes classes de langue d'année en année peuvent influencer la décision de choisir ou non cette langue en particulier. Via les termes d'interaction, on regarde aussi si ces influences varient selon la CSP des parents. Cependant, des changements de population en terme de classes sociales dans les différentes langues peuvent être dûs à de nombreux facteurs extérieurs inobservables, dont notamment des variations de population globale au sein du collège au cours des années, la création d'un autre collège à proximité, etc. Avec les variables dont nous disposons, nous décidons de réduire l'échantillon d'étude aux collèges au sein desquels la répartition globale de la population en terme de classe sociale n'a pas trop évolué.

4 Principaux résultats

4.1 Une répartition des classe sociales inégales à travers les langues

Le diagramme en bâtons ci-dessous (figure 3) regroupe toutes les données sur les élèves ayant étudié au collège entre 2005 et 2018. On y lit pour chaque LV2 et pour l'ensemble de nos données (barre TOTAL en haut en guise de référence), la proportion de ceux qui ont au moins un parent qui appartient à une CSP+. On retrouve ici les résultats de l'étude du CNESCO ayant motivé notre étude : il existe clairement une différence de répartition des enfants de CSP+ dans les langues. Tandis que la proportion des élèves avec un parent CSP+ qui font LV2 espagnol avoisine la proportion globale des enfants de CSP+, une langue comme l'allemand apparaît particulièrement attractive pour les enfants de parents CSP+. Au contraire, parmi les élèves qui font des LV2 plus rares comme le chinois, le japonais, le russe ou le portugais, la répartition est plus marquée et reste très spécifique à chaque langue. Ces résultats permettent de nous conforter dans l'idée de travailler sur la distinction LV2 allemand VS LV2 espagnol : tandis que l'une semble plutôt neutre vis-à-vis de la proportion d'enfants de CSP+ qui l'étudie, l'autre (l'allemand) semble davantage marquée socialement, ce qui invite à s'interroger sur l'origine de cette différence.

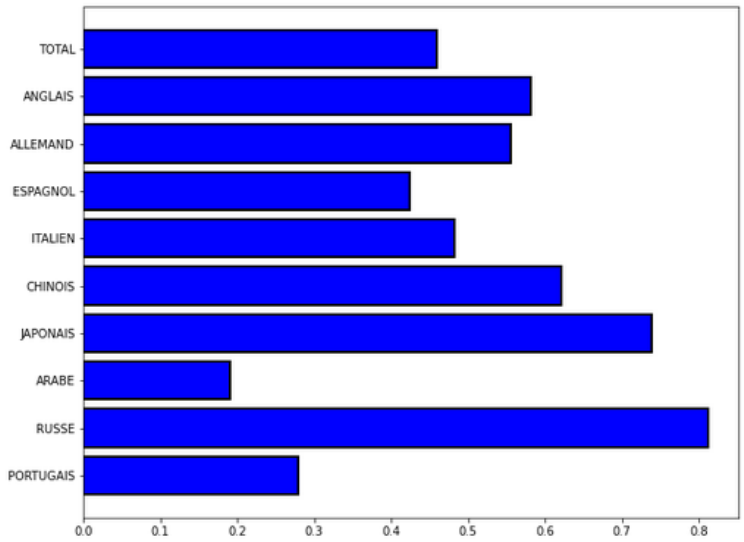

Figure 3: Proportion des élèves avec un parent CSP+ au total et dans les différentes LV2 pratiquées

4.2 L'évolution temporelle des CSP au sein des collèges et dans les langues

Pour justifier la pertinence de l'utilisation d'un indice de dissimilarité pour tester éventuellement un effet sur le choix des langues, nous nous sommes intéressés comme mentionné précédemment aux collèges au sein desquels les taux de CSP+ n'avaient pas grandement évolué au cours des années. Cependant, si on sélectionne au sein de ces mêmes collèges ceux où il y a toujours eu des élèves étudiant l'espagnol et l'allemand comme seconde langue, on peut regarder si les taux de CSP+ dans ces deux langues ont quant à eux évolué de manière plus visible. C'est ce qu'on observe sur la figure 4 : **l'abscisse** représente les environs 650 collèges retenus pour l'analyse. On trouve en **ordonnée** la valeur du plus grand écart de taux de CSP+ entre les années 2005 et 2018 à l'échelle globale du collège (en *vert*), dans la LV2 allemand (*bleu*) et LV2 espagnol (*rouge*).

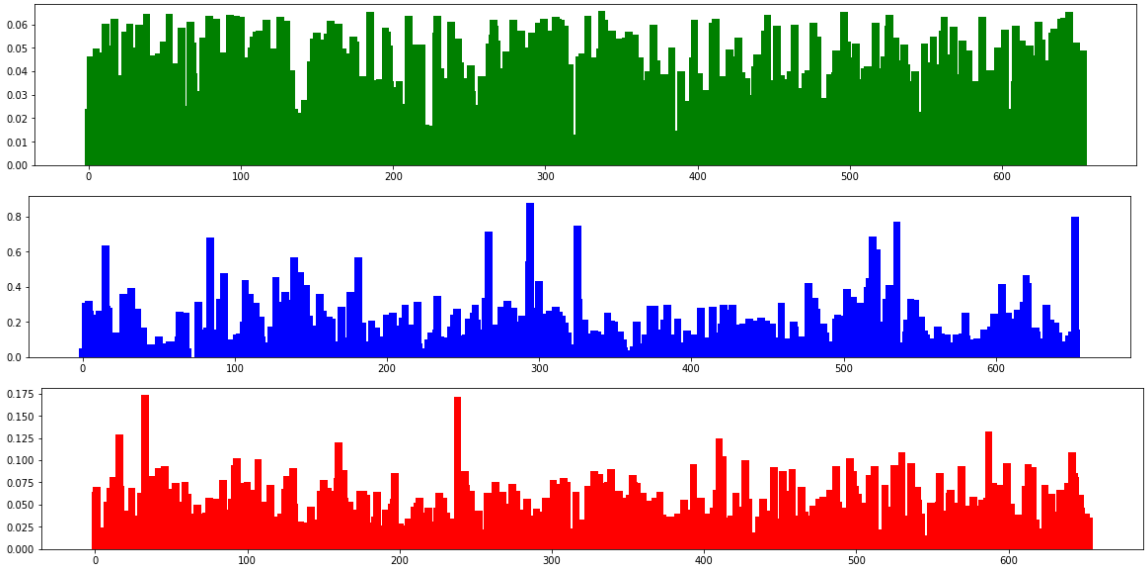

Figure 4: Répartition des plus grands écarts de taux de CSP+ entre les années à l'échelle globale des collèges, en allemand et en espagnol

Autrement dit, ces trois graphes représentent à quel point le taux de CSP+ a évolué entre les années

considérées au niveau global, en espagnol, et en allemand. Conformément aux collèges choisis, les valeurs en ordonnée vertes ne dépassent pas 0.06, mais pour les graphes rouges et bleus, c'est autre chose : dans certains collèges, le plus grand écart de taux de CSP+ entre des années atteint les 0.8 en allemand et presque 0.18 en espagnol. Cela signifie que même dans les collèges plutôt stables en terme de population CSP+, les classes de langue (et surtout celles d'allemand), ont tendance à voir leur population en terme de classes sociales évoluer. Ceci peut suggérer qu'il existe des interactions et des changements d'année en année au sein même des collèges, comme le suppose notre modèle.

4.3 Modèle de prédiction de la langue choisie

Nous avons donc réalisé sur Python la régression décrite dans la partie 3. Nous obtenons les résultats suivants :

Logit Regression Results

Dep. Variable:	lv2	No. Observations:	82482
Model:	Logit	Df Residuals:	82474
Method:	MLE	Df Model:	7
Date:	Tue, 10 May 2022	Pseudo R-squ.:	0.01225
Time:	21:31:21	Log-Likelihood:	-29395.
converged:	True	LL-Null:	-29759.
Covariance Type:	nonrobust	LLR p-value:	3.589e-153

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7968	0.096	-29.169	0.000	-2.985	-2.609
Résultat collège * CSP+	0.3347	0.179	1.868	0.062	-0.016	0.686
Dissimilarité * CSP+	1.7408	0.211	8.238	0.000	1.327	2.155
Dissimilarité	0.7986	0.175	4.554	0.000	0.455	1.142
tauxcspallp	-0.6976	0.091	-7.654	0.000	-0.876	-0.519
csp+?	0.1447	0.120	1.209	0.227	-0.090	0.379
SEXE	-0.0005	0.022	-0.022	0.983	-0.043	0.042
Résultats brevets année précédente	1.3202	0.177	7.450	0.000	0.973	1.668

Figure 5: Résultats de la régression sur l'année 2009

Le Pseudo R-carré est faible[1]. Cela signifie que nos variables ne permettent pas d'expliquer entièrement le modèle. Ce résultat est dû à de nombreuses variables non observées comme par exemple le "goût" pour une langue. Pour ce qui est des coefficients, nous trouvons des valeurs significatives pour lesquels nous pouvons donc bien interpréter le signe. Nous retrouvons les résultats attendus, décrits plus haut, à savoir : un effet de dissimilarité positif sur la probabilité de choisir allemand et des termes d'interaction positifs entre la CSP+ avec les résultats et la dissimilarité.

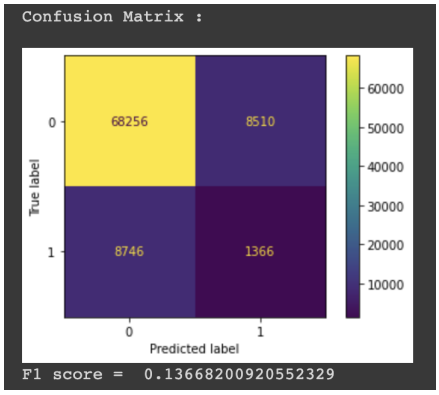

Figure 6: Résultats de la prédiction sur l'année 2014

[1]Une bonne valeur est comprise entre 0.2 et 0.4.

A partir de ce modèle de régression, nous avons cherché à savoir si il était possible de prédire le choix de la LV2 sur des donnés d'autres années. Sur la base de 2014, nous obtenons les résultats ci-dessus (figure 6). Nous obtenons un f1 score de 0.14. Ceci montre bien que notre modèle ne permet pas de définir les choix des élèves avec précision. Nous prédisons en trop grande majorité le choix de la langue espagnole.

5 Discussion

Pour le travail explicatif du choix de la CSP, nous aurions aimé avoir des résultats de brevet plus fin. Si nous avions à notre disposition les résultats du brevet séparés par la langue vivante choisie par les élèves, nous aurions pu comparer quel est effet est le plus significatif. C'est-à-dire que pour un élève avec une CSP donnée, nous aurions pu savoir si il préfère choisir sa LV2 pour être avec une CSP précise ou parce que celle-ci "promet" de meilleurs résultats. Dans cette étude, la variable de résultats des collèges au brevet n'a été utilisée que comme une variable de contrôle ou dans une variable d'interaction alors qu'elle aurait pu permettre d'expliquer le choix.
Dans ce travail, nous avons cherché à justifier le choix de la langue vivante comme un mécanisme de d'auto-ségrégation sociale, à savoir un choix qui permettrait aux classes sociales supérieures de se regrouper. Seulement, différents autres facteurs peuvent justifier le choix d'une langue. Il est possible qu'un élève ait une nationalité étrangère qui pourrait motiver son choix. Nous avons vu dans les statistiques descriptives que certaines langues étaient très marquées socialement, notamment représentante d'une population immigrée plus aisée ou non. Ceci peut alors être une conséquence directe de la surreprésentation des CSP+ dans certaines langues. De plus, il est également possible qu'un collégien ait une envie particulière motivée par un quelconque intérêt pour certaines cultures ou langues. Ces intérêts ne sont pas quantifiables et il n'est donc pas possible d'obtenir de telles informations. Ces variables explicatives auraient pu permettre d'obtenir un modèle plus approprié.

Conclusion

Dans cette étude, nous nous sommes appuyés sur des résultats descriptifs et qualitatifs montrant des inégalités dans les répartitions des CSP au sein des langues étudiées à l'école. Notre objectif était de montrer que le choix binaire entre l'allemand et l'espagnol pouvait être motivé par le fait que l'allemand attire en général plus d'élèves provenant de CSP supérieures. Pour ceci, nous nous sommes intéressés à l'indice de dissimilarité pour l'allemand dans les collèges. Si notre modèle nous a permis par exemple de montrer que l'effet de cette dissimilarité augmentait quand au moins un des parents appartenait à une CSP+, trop de variables semblent manquer à notre modèle pour pouvoir le considérer comme efficace et fiable. Le choix d'une langue reste plutôt personnel et expliqué par des facteurs souvent inobservables. Il est possible que ce soit pour ce type de raisons qu'aucune étude économétrique existent précisément sur le choix des langues au collège. Les études les plus avancées sur ces mécanismes d'auto-ségrégation sociale sont des études sociologiques où les auteurs ont mené des entretiens permettant de mieux cerner certains choix des parents pour l'école de leur enfant, ou s'arrêtent à des résultats descriptifs, comme l'étude du CNESCO.

References

[1] Offre de langues et choix : étude originale du cnesco. *CNESO*, 1.0, 2019.

[2] François Baluteau. Curriculum optionnel et composition sociale. le cas des collèges. *Socio-logos. Revue de l'association française de sociologie*, (8), 2013.

[3] Observatoire des inégalités. Les inégalités sociales sont fortes dès le primaire et le collège. *Khulna University Studies*, 1.0, 2019.

[4] Marco Oberti. *L'école dans la ville: ségrégation-mixité-carte scolaire*. Presses de Sciences Po, 2007.