1

MODÉLISER, POUR QUOI FAIRE ?

Michèle Tribalat
Septembre 1996

Question banale, mais qui mérite d'être posée pour lancer un vrai débat sur l'objet réel des modélisations. Le recours aux modèles a envahi les sciences sociales de deux manières. D'une part, les statisticiens se sont aventurés sur ce domaine avec leurs outils et leur savoir faire et, d'autre part, les chercheurs des sciences sociales ont importé des techniques nouvelles avec probablement moins de maîtrise. Mais, dans les deux cas, la technicité est apparue comme un critère d'auto-validation. Les régressions logistiques, notamment, ont connu un succès fulgurant, à tel point qu'aujourd'hui, il est mal séant de ne pas y recourir. Il semble cependant que cette technique, au demeurant très séduisante, puisqu'elle permet de prendre en compte conjointement plusieurs variables pour tenter d'expliquer au mieux un phénomène, n'a pas les vertus magiques qu'on lui suppose et mérite qu'on examine attentivement les conditions nécessaires à son application.

Ayant nous-mêmes succombé à l'effet de mode, nous avons cherché, en recourant à cette technique statistique, à expliquer les facteurs de réussite au BAC en France, en nous penchant plus précisément sur les spécificités éventuelles des jeunes d'origine étrangère. La scolarité est un sujet très investi par la statistique, car le ministère de l'Éducation nationale dispose de données exceptionnelles à travers le suivi de panels d'élèves, fruit de l'activité de la Direction de l'évaluation et de la prospective. De nombreuses études statistiques ont déjà été effectuées portant sur un, voire deux cycles d'enseignement, mais sans intégrer l'aboutissement du parcours secondaire[1]. Il nous a donc semblé intéressant de mener une analyse multivariée de l'ensemble du parcours à partir d'une enquête rétrospective réalisée par l'INED[2]. Cette étude a été menée avec une préoccupation d'analyste, plus que de statisticien, ce que nous ne sommes pas. En béotienne, nous avons pris au sérieux les hypothèses de fonctionnement des régressions logistiques, ce qui nous a conduit, en dépit de notre bonne volonté première, à une remise en cause fondamentale des pratiques en matière de modélisation. Quel est le sens réel des variables incluses ? Les exigences constitutives des modèles additifs sont-elles compatibles avec la nature des variables choisies comme explicatives ?

---

[1] Les plus récentes sont celles mises en oeuvre par Louis-André Vallet et Jean-Paul Caille. "Les élèves étrangers ou issus de l'immigration dans l'école et le collège français", Les dossiers d'éducation et formations, Avril 1996, n° 67.

[2]Des résultats synthétiques de ce travail ont été publiés. Cf. M. Tribalat, "La réussite au Bac des jeunes d'origine étrangère, in : À l'école de la République, Hommes & Migrations, N° 1201, septembre 1996.

Nous avons ainsi été amenée à distinguer les valeurs d'ajustement d'un modèle de sa puissance explicative et, par là même, à nous poser la question d'une éventuelle distorsion entre "comprendre" et "prévoir".

Notre réflexion s'appuie donc sur une expérience pratique de modélisation appliquée à la recherche des déterminants de la réussite scolaire et notamment celle des jeunes d'origine étrangère, à laquelle il sera fait allusion, à titre d'illustration.

Quelque soit le domaine d'étude, il est très rare de disposer d'effectifs suffisants pour croiser (en arborescence) l'ensemble des variables dont on suppose qu'elles agissent sur le phénomène qui nous intéresse. De ce point de vue, les enquêtes spécifiques contiennent de nombreuses informations collectées à dessein pour permettre des analyses fines. Mais, les croisements simples de variables, même judicieux, ne permettent pas toujours de cerner l'agencement de leur intervention, du fait de la limite des effectifs enquêtés. L'idée de faire intervenir conjointement les différentes variables à travers un modèle, permettant d'éliminer les effets de structure résultant de corrélations entre ces variables est donc extrêmement séduisante. Il s'agit alors d'expliquer une variable A par des variables Xi, supposées intervenir de manière additive, ce qui exige une absence d'interactions entre elles.

Une première évidence ressortant de notre étude réside dans la rareté des variables indépendantes. Compte tenu de la difficulté de réunir, en général, les conditions satisfaisantes pour parler d'*effet propre* (absence d'interaction), il semble préférable de substituer à ce terme, avant d'avoir tranché, c'est-à-dire au stade de la modélisation sans interaction, celui d'*effet net*.

Faut-il rejeter, avant tout examen plus approfondi, les variables d'effet net nul ?

Si on peut conclure à l'absence d'effet propre en cas de nullité des coefficients (voisins de zéro) affectés à l'ensemble des modalités d'une variable X1, un tel constat ne suffit pas pour rejeter cette variable, plutôt qu'une autre. Il est nécessaire de s'assurer qu'elle n'entre pas en interaction avec d'autres variables du modèle et qu'il n'existe pas certains segments de populations réactifs à la variable X1. L'absence d'effet net peut cacher, en fait, des compensations d'effets contradictoires.

Par exemple, on n'observe pas, globalement, de différence significative de réussite scolaire entre les jeunes d'origine française[3] et ceux qui ne le sont pas (effet net nul de l'origine ethnique). Mais, les jeunes d'origine française réussissent mieux que les autres lorsqu'ils appartiennent à des familles de moins de 5 enfants. Au contraire, les performances des enfants d'origine étrangère sont meilleures que celles des enfants d'origine française des familles nombreuses. Ainsi, l'équivalence globale de

[3] Nés en France de parents nés en France

performances entre jeunes d'origine française et jeunes d'origine étrangère résulte de la compensation d'un avantage des premiers dans les petites familles et des seconds dans les familles nombreuses.

Compte tenu du fait que l'absence d'effet propre est plus la règle que l'exception, on ne voit pas pourquoi on éliminerait d'un modèle des variables au motif que la compensation des effets contradictoires est totale. Lorsqu'on examine la question des interactions, toutes les variables doivent donc être testées, y compris celles dont l'effet net est nul.

Validation de l'hypothèse d'une absence d'interaction

Il s'agit là d'une hypothèse classique en démographie nécessaire à la mesure d'un phénomène en fonction du temps, lorsqu'un autre phénomène entraîne des sorties d'observations. Pour la rigueur de la mesure, il est nécessaire que ces sorties d'observations ne soient pas sélectives par rapport au phénomène étudié. Souvent, nous sommes confrontés à la difficulté de prouver la validité de l'hypothèse nécessaire à la rigueur de la mesure. Lorsque les sorties d'observation ne peuvent être négligées et que l'hypothèse d'une sélection est prouvée, ou seulement hautement vraisemblable, il est recommandé alors de scinder le champ d'observation en sous-champs correspondant au fait d'avoir subi ou non l'événement qui fait sortir du champ d'observation.

Dans le cas d'une modélisation, la question de l'interaction se pose un peu différemment et se trouve, en quelque sorte, démultipliée. L'hypothèse d'une absence d'interaction entre les variables explicatives est nécessaire pour valider le caractère additif du modèle et mettre ainsi en évidence les effets propres de ces variables.

Il faut bien distinguer deux sortes de liaisons : corrélation simple et interaction.

Le premier type de liaison n'est pas contre-indiqué et, d'une certaine manière, justifie le recours à un modèle : c'est bien parce qu'on suppose des liaisons entre variables que l'examen séparé des variables ne suffit pas et qu'on utilise une régression multiple. On cherche par là à éliminer les effets de structure et à s'affranchir des corrélations entre les variables explicatives.

Par exemple, si l'on suppose que la taille de la fratrie et l'origine sociale peuvent influer sur le parcours scolaire des jeunes, il faut pouvoir tenir compte du fait que la taille des familles est corrélée à l'origine sociale : on trouve plus de familles nombreuses dans les familles ouvrières que dans les familles des classes moyennes et supérieures.

Une interaction entre deux variables X1 et X2 suppose que l'effet global observé d'une variable X1 sur un phénomène A va changer significativement (s'annuler ou être de sens contraire) sur certains segments de populations découpés par les modalités de la variable X2. Dans ce cas, on ne pourra parler d'effet propre de la

variable X1. Si, comme c'est souvent le cas alors, la variable X2 change aussi d'effet en fonction des modalités de X1, elle sera également déclarée sans effet propre. Il suffit d'une variable Xi changeant l'effet de la variable X1 pour conclure à l'absence d'effet propre de cette dernière. Tester les interactions revient généralement à croiser terme à terme les modalités des variables. Une telle démarche rencontre assez souvent des limites liées aux effectifs disponibles et à la polarisation des caractéristiques sur certaines modalités.

On peut donc trouver des variables non corrélées, mais entrant en interactions entre elles. C'est par exemple le cas du sexe et de l'origine ethnique : les familles d'origine française n'ont pas tendance à avoir plus de filles ou plus de garçons que les autres, mais les filles d'origine étrangère réussissent mieux que les garçons, alors que les performances des filles d'origine française sont équivalentes à celles des garçons.

On pourra également avoir deux variables explicatives qui, tout à la fois, sont corrélées et entrent en interaction. Par exemple, les jeunes d'origine étrangère dans leur ensemble[4] redoublent plus souvent dans le primaire lorsqu'ils appartiennent à une famille nombreuse. Par ailleurs, le fait d'avoir redoublé au moins une classe du primaire ne nuira pas plus aux enfants des petites familles qu'à ceux des grandes, alors qu'un avantage se dégage à la faveur des familles réduites en l'absence de redoublement. Autrement dit, il n'y a pas forcément cumul de handicaps : si globalement, l'appartenance à une famille nombreuse s'avère un désavantage, il ne s'additionne pas à celui lié au redoublement dans le primaire.

Tout bien réfléchi, il faut la foi du charbonnier pour croire un seul instant que l'absence d'interaction sera la situation la plus probable, car il n'existe pas de différence de nature entre variable expliquée et variables explicatives. La plupart du temps, on cherche à expliquer un fait ou une situation par d'autres faits, d'autres situations.

Comment expliquer le faible intérêt des utilisateurs des modèles de type logit pour la validation de l'hypothèse d'absence d'interaction ?

La plupart du temps, les auteurs qui développent ce type de modèles sont muets sur l'éventualité d'une impasse liée à l'existence d'interactions multiples. Indépendamment de l'optimisme irréductible de certains "modélisateurs" qui se traduit par un manque de curiosité pour la validation du modèle, la croyance en l'absence d'interactions provient souvent d'une confusion entre qualité d'ajustement du modèle et pourvoir explicatif de ce dernier[5]. Ajoutons à cela une certaine exigence sur le niveau

[4] jeunes des trois origines suivantes pris conjointement : algérienne, espagnole ou portugaise.

[5] Le plus souvent, en effet, on risque de se limiter à la comparaison globale du $\kappa 2$ obtenu sans interaction de la variable X1 avec la variable X2, entrées de manière additive d'une part et le $\kappa 2$ de l'ajustement

des seuils de significativité et on s'explique facilement pourquoi la question des interactions est généralement passée sous silence. Dans le domaine des performances scolaires, nous avons trouvé les interactions les plus intéressantes, en termes de sens, à des seuils que tout bon statisticien aurait rejetés.

Prenons un exemple, celui de la combinaison de l'âge d'entrée à l'école avec l'origine sociale dans l'explication des performances au BAC des jeunes d'origine française. L'entrée séparée des variables montre un avantage à l'appartenance au haut de la hiérarchie sociale et à l'entrée à l'école précoce. L'interaction indique la réduction totale de l'avantage social, au moins pour les enfants d'indépendants (artisans, commerçants, chef d'entreprise) en cas de scolarisation tardive et l'absence d'influence de l'âge d'entrée à l'école sur la scolarité des enfants d'origine ouvrière. Ainsi, ni l'origine sociale, ni l'âge de scolarisation n'ont d'effet propre. Et pourtant la comparaison des ajustements avec ou sans interaction fait ressortir un seuil de significativité global de 36%, inacceptable pour n'importe quel statisticien normalement constitué.

En outre on a pu constater que les interactions étaient plutôt le lot commun que l'exception. En France, dans l'ensemble des jeunes nés en 1963-71, une seule des 7 variables incluses dans le modèle exerce un effet propre sur la scolarité, celle que nous avons appelée "effet famille", explicitée à travers le fait d'avoir ou non au moins un frère ou une soeur dans l'enseignement supérieur. Sans décrire une qualité intrinsèquement supérieure de certaines familles, cette variable reflète plutôt l'engagement des parents et de la famille toute entière (ce peut être l'aide des frères et sœurs par exemple) par rapport à l'école ; c'est pourquoi on parle alors d'un effet famille. La combinaison de quelques variables entre elles ne permet pas de sortir de l'excès de complexité mis à jour.

Réduit-on la complexité et l'enchevêtrement des variables en découpant le champ ?

Lorsqu'une variable "explicative" entre en interaction avec de nombreuses autres, on peut être tenté de découper le champ de modélisation en fonction de cette variable.

Dans notre étude sur les performances scolaires, nous avons essayé de gagner en clarté en découpant le champ en fonction de l'origine ethnique, tant les interactions avec cette variable sont nombreuses. Ainsi, nous avons considéré globalement, puis

---

intégrant un croisement des modalités des deux variables d'autre part. Si l'ajustement ne semble pas significativement amélioré par la combinaison, on conclura à l'absence d'interaction. Or, le $\kappa 2$ global caractérisant une variable, ou une combinaison de variables dépend beaucoup du regroupement des modalités effectué. Par ailleurs, une faible significativité globale n'empêche pas l'existence d'un effet particulièrement significatif de l'une de ces modalités.

séparément, les trois groupes de jeunes d'origine étrangère enquêtés : ceux d'origine algérienne, espagnole et portugaise.

Or, le schéma des interactions n'a pas tendance à se simplifier et, en tout cas, ne se répète pas à l'identique d'un groupe à l'autre, lorsqu'on passe de l'ensemble des jeunes, à chacune des origines ethniques prises séparément. Certaines variables jouent plus ou moins et sur des segments de population non identiques d'un groupe à l'autre. Ainsi, l'âge de scolarisation, sans effet net, au niveau global, lorsqu'on considère conjointement les trois groupes de jeunes d'origine étrangère, s'avère intervenir sur la scolarité des jeunes d'origine portugaise, l'effet étant alors contraire à celui relevé sur les Français de souche, avec une "prime" à la scolarisation tardive.

D'autres variables, agissant d'une manière relativement simple, se révèlent avoir des actions plus complexes lorsqu'on découpe le champ par origine ethnique. Autrement dit, il ne semble pas y avoir d'intervention forcément logique et univoque des variables et le découpage du champ ouvre sur des complexités nouvelles. On pense aux fractales, à ceci près que les arborescences successives ne se ressemblent pas.

Quel sens finalement attribuer aux variables ?

Tout bien considéré, il faut un certain aveuglement pour espérer expliquer une réalité complexe à partir de faits dont on ne voit pas, a priori, quel sens leur donner.

Trois exemples permettent d'illustrer cette réflexion : de l'influence du sexe, de l'origine ethnique et de la taille de la famille, sur la réussite au BAC.

- Toutes les modélisations sur la réussite scolaire intègrent la variable sexe qui aboutit, aujourd'hui, à mettre en valeur la meilleure réussite des filles. Quel sens donner à un tel résultat ? Doit-on en déduire que la constitution génétique des filles les prédisposerait à des meilleures performances ? Dès qu'on se pose la question du sens, on arrive vite à l'absurdité du contenu. En fait, cet avantage net n'est que le solde d'effets multiples qui se compensent imparfaitement mais, au total, à l'avantage des filles ! La mise en oeuvre des interactions entre le sexe et d'autres variables du modèle permet vite de s'en convaincre.

- La question principale abordée à travers les tests statistiques sur la réussite scolaire des jeunes d'origine étrangère (le plus souvent des enfants étrangers) vise à déterminer l'existence d'un effet lié à l'origine ethnique, origine bien souvent réduite à la seule nationalité : les enfants français réussissent-ils mieux que les enfants étrangers ?

  Là encore, il est nécessaire de se demander, a priori, quel est le sens d'une telle interrogation. Elle suppose qu'on imagine comme vraisemblable qu'une simple

appartenance ethnique puisse constituer, en soi, un avantage ou un handicap. Il y aurait ainsi, comme pour le sexe, un avantage possible à être né, non plus fille au lieu de garçon, mais ici plutôt qu'ailleurs. D'une certaine manière, on essentialise ainsi l'origine ethnique (le plus souvent la nationalité). L'étonnement et le contentement médiatique accompagnant l'annonce d'une absence de différences (voire même d'un avantage lié au fait d'être étranger), une fois qu'on égalise les conditions de la comparaison, surprend[6]. En effet, ils dénotent un soulagement qui en dit long sur les représentations en termes de hiérarchisation ethnique qui traversent la société française. Mais ce contentement porte aussi le risque de faire croire, au nom de cette absence de différences, qu'il n'y a rien à faire, en termes de politique spécifique, pour les jeunes d'origine étrangère, ou certains d'entre eux ; ce qui est peut-être vrai, mais mérite d'être vérifié. Il peut faire croire aussi que l'institution scolaire n'est pas traversée par des pratiques discriminatoires, ce qui reste également à démontrer.

L'absence de significativité globale de la variable ethnique en France (entre Français de souche et non Français de souche) résulte de compensations d'effets entre différents segments de populations découpés par les autres variables du modèle. La variable ethnique, sans valeur intrinsèque, entre en interaction avec pratiquement toutes les variables du modèle. Elle n'a pas d'effet propre sur la scolarité, sans impliquer une identité des autres caractéristiques. Au contraire, cette variable médiatise des normes et représentations différentes, ce qui incite à la plus grande prudence quant à l'établissement de règles "universelles" sur les attributs de la réussite scolaire.

- Dans la plupart des études, une petite taille familiale a toujours été associée à une meilleure réussite scolaire. Même si, dans ce cas, on peut trouver un sens à ce résultat (plus grande disponibilité des parents, signe d'une ambition sociale plus élevée, dans la mesure où une petite famille permettra aux parents d'assumer les frais d'une scolarité longue ...), la question qui se pose est de savoir si cet effet est indépendant des normes en matière de taille de famille. Il ne l'est pas, comme l'indique l'interaction avec l'origine ethnique même si les normes familiales favorisant les familles nombreuses constituent un handicap en soi.

  Par exemple, pour les jeunes d'origine algérienne, vivre dans une petite famille (de deux ou trois enfants) est une aberration qui conduit aux résultats les plus mauvais en matière de scolarité. Les meilleurs résultats sont obtenus dans les familles comprenant au moins cinq enfants. Au contraire, pour les jeunes d'origine

[6] On fait référence ici à l'écho médiatique et scientifique qui a accompagné la sortie des résultats de l'étude réalisée à partir d'un panel d'élèves suivi par la DEP. Cf L-A Vallet et J-P Caille, op. cit.

espagnole ou portugaise, une taille familiale plus restreinte est plus propice aux bons résultats scolaires. Ceci dit, les enfants d'origine algérienne appartenant à une famille de cinq enfants ou plus ne font pas mieux que ceux d'une autre origine vivant dans une famille de taille identique. Or, la taille moyenne des familles dans lesquelles vivent les jeunes d'origine étrangère est la plus élevée chez les jeunes d'origine algérienne (7,6 enfants) et la plus faible chez les jeunes d'origine espagnole (3,9 enfants). Elle atteint 4,4 enfants chez les jeunes d'origine portugaise. L'effet de la taille de la famille sur la réussite scolaire dépend donc fortement des normes en la matière.

En réalité, les faits, les états entrant dans les modèles, ne sont généralement pas d'un ordre différent de celui qu'on cherche à expliquer et n'ont donc pas ou peu de force explicative. Tous reflètent (au sens littéral du terme) des normes et des représentations, qui rarement se laissent emprisonner à travers un fait, ou l'état d'une personne.

La sagesse voudrait alors qu'on ait le courage d'avouer que, sauf exception, les variables dites explicatives n'ont pratiquement aucun pouvoir explicatif réel, mais reflètent simplement la condensation d'effets venus d'ailleurs. On s'est trompé, on n'a pas mis les bonnes variables dans le modèle. La plupart des enquêtes statistiques, consacrées avant tout à la mesure et la description de phénomènes, n'ont pas vocation à cerner les normes et représentations qui "travaillent" la société française. Ces "insuffisances" invitent au doute et à la modestie. Elles incitent aussi à ne pas fonder une confiance aveugle dans les modèles factuels et à se méfier des automatismes "panurgiens".

Comprendre ou/ et prévoir et la nécessaire simplification de la complexité

Les réflexions conduites à partir d'un modeste travail de modélisation des facteurs de réussite au BAC montrent très clairement que la compréhension de la réalité ne passe pas par la prise en compte d'une multiplicité de caractéristiques individuelles. Tout au plus font elles surgir des hypothèses, au demeurant fort incomplètes, sur les normes et les représentations déterminant les comportements.

En fait, cet exemple montre bien que, plus on multiplie l'information individuelle, plus on la découpe, moins elle est lisible et plus elle semble ouvrir sur une complexité toujours plus grande.

Vouloir saisir l'individu dans toute sa diversité et entrer le plus finement dans la complexité des comportements individuels, dans l'espoir d'y trouver une cohérence cachée, illisible autrement, laquelle ouvrirait des perspectives infinies sur la compréhension de la société, semble être la tendance actuelle vers laquelle se dirige la

recherche en sciences sociales. Il y a pourtant une contradiction inhérente à cette démarche visant à mettre en évidence des singularités, allant presque jusqu'à à la singularité individuelle, pour espérer aboutir à une meilleure compréhension globale. On se fourvoie si l'on espère faire fructifier notre connaissance "micro", visant à l'étude fine des comportements, pour en déduire, par additivité en quelque sorte, des relations "macro", outils d'une meilleure prévision. C'est nier, par là, l'existence d'entités collectives, chères à Élias Canetti[7], lesquelles entités ne se résument pas à l'addition pure et simple d'individus, mais s'expriment à travers des normes et représentations.

Tant qu'on ne saura pas faire autre chose que chercher les déterminants de situations à travers des myriades d'états individuels, n'apportant que peu d'éléments de compréhension réelle, il paraît plus sage de chercher à prévoir à partir de relations simples et stables, en laissant de côté ce pseudo savoir, avec, bien sûr, l'inconvénient de ne pouvoir anticiper les discontinuités et retournements de tendance, ce qui est d'ailleurs généralement le cas actuellement des exercices de "prévision".

La tarte à la crème d'un monde complexe et de la nécessité de penser la complexité masque, beaucoup plus simplement, un manque de réflexion sur les limites de la description indéfiniment découpée, mais stérile, par rapport à l'intelligence analytique.

[7] Fasciné dans sa jeunesse par une manifestation à Vienne, il s'est beaucoup intéressé au concept de foule.