Questions de méthode

Modéliser l'insertion

par Jean-Pierre Fénelon, Yvette Grelet et Yvette Houzel*

Les débats autour des outils d'analyse des trajectoires individuelles ont souvent renvoyé dos à dos démarche typologique et défenseurs des modèles économétriques. En examinant un ensemble d'études récentes, les auteurs nous montrent que la sélection et la mise en forme des données et le choix de la grille d'interprétation sont aussi déterminants que les options de méthode. Les deux démarches sont d'ailleurs plus complémentaires qu'antagoniques[1].

Ces dernières années ont vu se multiplier en France comme à l'étranger les études fondées sur l'exploitation de données longitudinales, tant pour conduire des travaux d'évaluation des dispositifs d'aide à l'emploi, que pour analyser les processus d'insertion des jeunes. Une telle profusion de résultats, s'appuyant sur des méthodes diverses et jugées souvent complexes, conduit de plus en plus à s'interroger sur leur appropriation par les acteurs des politiques publiques [2]. Le fait que ces travaux aboutissent parfois à des conclusions contradictoires, au moins en apparence, vient encore renforcer la perplexité des destinataires ou utilisateurs de ces recherches.

Pour tenter de répondre à ce besoin de clarification, nous nous sommes livrés à une comparaison d'une trentaine d'études françaises, présentées par des économistes, des sociologues ou des statisticiens, au cours des « Journées du Longitudinal » [3]. Nous nous sommes attachés à dérouler l'ensemble de la chaîne d'analyse commune à toutes les études, laissant de côté un grand nombre d'aspects qui ont déjà été abordés dans des panoramas plus généraux de l'évaluation des politiques publiques [4]. Nous avons seulement tenté de repérer à chaque étape les choix statistiques qui y sont faits et qui peuvent être source de divergence. Afin de faciliter les comparaisons nous avons sélectionné les travaux s'appuyant sur des enquêtes rétrospectives standardisées (telles les enquêtes de cheminement du Céreq) : la comparabilité des études n'est pas assurée pour autant, en raison du manque de concepts communs et d'un vocabulaire harmonisé.

Les différences de méthode contribuent aussi à ces divergences : dans les études rencontrées, il s'agit surtout de modélisation économétrique ou d'analyse typologique. Leur choix renvoie à des questions épistémologiques complexes : on sait les débats récurrents que soulève l'analyse de la nature des relations entre la statistique et les sciences sociales – économie

* **Jean-Pierre Fénelon** est directeur de recherche au CNRS, Laboratoire d'économie sociale (URA n° 941). Statisticien spécialiste des relations entre l'informatique et la statistique chez IBM, puis directeur de recherche au CREDOC, il a co-animé avec Michel Volle le séminaire « Analyse des données et Économie ». Aujourd'hui, il travaille sur les rapports entre ces deux disciplines, notamment sur les aspects décisionnels de l'Analyse des données. Il a publié *Statistique et informatique appliquées*, Lebart L., Fénelon J.-P., Dunod, Paris, 1975 et *Qu'est-ce que l'Analyse des Données*, Fénelon J.-P., Lefonen, Paris, 1981.

Yvette Grelet est chargée d'études au Département des entrées dans la vie active du Céreq. Ses travaux portent sur les déterminants structurels de l'insertion des jeunes, et plus récemment sur la contextualisation locale des débuts de parcours professionnels. Elle est l'auteur de : « Niveau, spécialité et région : des facteurs clés de l'insertion professionnelle », in *L'insertion professionnelle des jeunes, analyses et débats*, M. Vernières Dir., Economica 1997.

Yvette Houzel est maître de conférences à l'Université de Paris 1 au Laboratoire d'économie sociale (URA n° 941). Elle participe aux travaux sur l'insertion menés dans le cadre du CIA-Céreq -LES, et s'intéresse en particulier à l'analyse des calendriers des enquêtes Céreq en utilisant des méthodes d'analyse textuelle.
Elle a présenté, avec M. Le Vaillant, « Les trajectoires d'insertion professionnelle, une analyse de calendriers », aux XV[e] journées de l'Association d'économie sociale, Nancy, septembre 1995.

[1] Une version préliminaire de ce texte a fait l'objet d'une communication aux 4èmes Journées du Longitudinal, Paris, 1997. Ce travail a bénéficié des commentaires de Bernard Girard (SAMOS) et de Saïd Hanchane (Céreq). Qu'ils en soient ici remerciés.

[2] (Elbaum, 1997), (Fargher, 1997).

[3] Ces Journées sont organisées chaque année depuis 1994 conjointement par le Céreq et ses centres associés, et l'Institut du Longitudinal. Les actes en sont publiés dans la série des documents du Céreq cités en référence bibliographique.

[4] (Aucouturier, 1994), (Gautié, 1996), (Lechêne et Magnac, 1996), (Schömann, 1996).

ou sociologie [5]. Il n'est pas dans notre propos d'y contribuer, mais simplement de montrer que dans un domaine encore mal théorisé comme celui de l'insertion professionnelle, des oppositions affichées, comme « exploratoire / confirmatoire » ou « descriptif / explicatif » ont peu de fondement réel, et tiennent plus à l'habitude. La question d'établir quelle serait la méthode *la plus appropriée* au traitement des données longitudinales, qui, en France, se focalise principalement autour des avantages respectifs des exploitations fondées sur une modélisation de type économétrique et celles utilisant les méthodes factorielles ou typologiques, nous paraît ainsi manquer de pertinence.

Nous nous sommes fixés une grille de lecture simple. Après avoir posé le cadre de l'analyse et quelques définitions de vocabulaire, nous examinons les différents procédés de mise en forme des données (comprenant le choix des nomenclatures, la définition des trajectoires et le mode d'intégration du temps). Nous nous intéressons ensuite aux méthodes statistiques proprement dites, dont la place est précisée dans ce que nous esquissons comme un « schéma général de l'évaluation de l'insertion ». Enfin, nous proposons quelques pistes qui permettraient de dépasser les difficultés d'analyse soulevées par la complexification croissante des processus d'insertion et d'en enrichir la compréhension.

UN ESSAI D'HARMONISATION DU LANGAGE

Nous allons tenter de décomposer les différentes étapes intervenant dans l'analyse des *trajectoires* individuelles, après avoir posé quelques définitions (dont celle de trajectoire) qui fixeront le vocabulaire pour le temps de cet exposé.

Les données individuelles recueillies par les enquêtes de cheminement du Céreq permettent de décrire très finement la position d'un individu à chaque instant d'une période d'observation, sa situation instantanée. On peut utiliser pour reconstituer ces situations, non seulement le calendrier récapitulatif accompagnant le questionnaire [6], mais aussi toutes les informations collectées par le questionnaire lui-même et portant principalement sur les caractéristiques des emplois occupés. En fonction des objectifs de l'étude, on sélectionne les informations par lesquelles on veut décrire les parcours, et à partir desquelles on construit une *nomenclature d'états* élémentaires exclusifs : à chaque instant l'individu occupe une position et une seule, étant entendu qu'un état de la nomenclature peut très bien combiner des activités simultanées (« en études à temps plein exclusivement » / « en études et en emploi » / « en emploi exclusivement » / ...)

Tableau 1
Le Calendrier d'états
Exemple de l'enquête auprès des jeunes sortis de l'enseignement secondaire général ou technique en 1986

	1985			1986	1987	1988	1989		
	octobre	novembre	décembre	...	...	...	...	novembre	décembre
En emploi	01*	02	03					54	55
Au service national	01	02	03					54	55
En formation ou stage	01	02	03					54	55
En études à temps plein	01	02	03					54	55
Au chômage	01	02	03					54	55
Inactif	01	02	03					54	55
Naissance des enfants	01	02	03					54	55

* Numéro des mois

[5] Voir en particulier (Desrosières, 1996).

[6] Voir le tableau 1.

Nous appellerons *calendrier d'états* ce cadre vierge (l'enchaînement des mois, l'éventail des situations) pour le distinguer du *calendrier individuel* qui est le cadre rempli, la donnée de base de la trajectoire individuelle (les situations occupées mois par mois par un individu).

La *trajectoire* est le résultat d'une mise en forme du calendrier individuel. Lorsque deux auteurs parlent de « la trajectoire » d'un jeune, il est peu probable que ce vocable recouvre le même contenu. À supposer qu'ils partent de la même donnée de base, avec la même nomenclature d'états, ils divergeront sur l'opération de mise en forme : construction d'un ou plusieurs indicateurs, découpage du calendrier en périodes, arrangement de formes syntaxiques, calcul d'une matrice de transition, etc., autant d'images, autant d'interprétations différentes de la même donnée de base.

On parlera ici de la *chaîne d'analyse statistique* des trajectoires individuelles qui va de la mise en forme des données à l'interprétation des résultats. Nous ne discuterons pas du cadre analytique de l'insertion dans lequel les trajectoires font l'objet d'un traitement statistique en tant que données quantitatives, individuelles, collectées par enquêtes rétrospectives. Dans tous les cas cette chaîne est mise en place après qu'ont été fixés les objectifs, qui guident tous les choix intervenant dans son déroulement. La définition des objectifs relève de l'expertise scientifique du domaine (l'insertion des jeunes, le chômage,...). Elle s'inscrit dans des problématiques diversifiées (théories du capital humain, de la segmentation, des réseaux sociaux...). Même lorsqu'on ne formule pas explicitement d'hypothèse, on a malgré tout opéré une sélection dans les données de base (les individus, les variables et leur nomenclature), et cette sélection est porteuse de sens. Les données ont été collectées pour répondre à un ensemble précis de questions, et portent la trace des hypothèses qu'elles sont chargées de mettre à l'épreuve, et qui ont orienté la détermination du champ de l'enquête et du plan de sondage, aussi bien que la conception du questionnaire, la formulation des questions et leur codification.

Dans le corpus analysé, les objectifs poursuivis relèvent de l'évaluation. En effet, si la visée évaluative n'est pas toujours apparente dans les textes que nous avons passés en revue, elle n'en est cependant jamais absente. Au bout du compte, il s'agit toujours de hiérarchiser les parcours professionnels, et de rechercher parmi les caractéristiques individuelles, les formations initiales, les dispositifs d'aide à l'emploi, etc., ce qui explique les parcours professionnels « réussis ». Dans le cas de la démarche typologique, le problème essentiel est celui de l'identification des variables de résultat, dans le cas de la modélisation, la quantification des effets. Ce souci d'opérationnalité lié à l'évaluation explique probablement que la démarche économétrique soit le plus souvent de type empirico-déductif et que l'analyse des données se focalise sur des typologies.

Dans la chaîne d'analyse on peut distinguer plusieurs étapes :

- la définition des variables sélectionnées pour l'étude : choix des nomenclatures, construction d'indicateurs et de variables composites ;

- la mise en œuvre statistique proprement dite (« on fait tourner le programme ») ;

- l'interprétation et la formulation des résultats.

LA MISE EN FORME DES DONNÉES

UNE ÉTAPE DÉTERMINANTE : LA DÉFINITION DES NOMENCLATURES

La nomenclature des états

La trajectoire est définie à partir d'une nomenclature d'états, dont les frontières sont supposées étanches et sans ambiguïté (par exemple le chômage est clairement distingué de l'inactivité) ; et un découpage du calendrier en unités de temps pendant lesquelles l'individu observé occupe un état et un seul. La nomenclature constitue une partition de l'ensemble de toutes les situations possibles.

Le choix de la nomenclature des états est à notre avis parmi ceux qui ont le plus d'incidence sur la structure des données ; en dépendent les définitions de la trajectoire et des variables « à expliquer » (les durées *dans un état*, les transitions *entre états*, etc.). Ce choix est très lié au cadre théorique d'analyse et aux objectifs, qui conduiront par exemple à ne retenir que les descripteurs de l'emploi ; à en assimiler tous les statuts quels qu'ils soient, ou à regrouper CDI [7] et CDD [8] dans « l'emploi ordinaire », ou bien le CDD et les mesures de politique pour l'emploi dans l'« emploi précaire » ; à distinguer les mesures non marchandes des mesures marchandes [9], ou encore à assimiler

[7] Contrat de travail à durée indéterminée.

[8] Contrat de travail à durée déterminée dont l'utilisation est, en principe, limitée.

[9] Par exemple, les contrats emploi-solidarité sont des mesures non marchandes car réservées aux administrations, collectivités locales et associations alors que les contrats de qualification qui comportent une formation en alternance sont utilisés dans le secteur marchand.

l'apprentissage à ces dernières...autant d'options qui pèseront sur les résultats. Même en l'absence de référence à un cadre théorique, ces choix s'inscrivent dans une représentation du fonctionnement du marché du travail.

Cette nomenclature suppose d'entrer dans le cadre d'analyse de la relation formation-emploi qui a présidé à l'établissement du calendrier : on s'aperçoit, par exemple, à travers les changements de descriptif d'emploi du calendrier Céreq, que cette conception a changé avec les évolutions du marché du travail (tableau 2).

C'est en 1983, avec l'enquête de cheminement auprès des jeunes sortis d'apprentissage en 1978, que les calendriers ont été introduits au Céreq pour la première fois. Le questionnaire lui-même ne reconstituant que les périodes d'emploi, le calendrier avait l'avantage de combler à moindre frais le manque d'information sur les périodes intermédiaires, donc aussi sur un chômage juvénile de plus en plus préoccupant. De plus, le calendrier s'est révélé utile comme support d'interrogation, et pour contrôler la cohérence des informations fournies dans le questionnaire.

Au début des années quatre-vingt, c'est l'emploi salarié à temps plein qui constitue la norme d'emploi. Dans la première version du calendrier, les diverses situations d'emploi salarié sont indifférenciées, qu'il s'agisse de contrats à durée déterminée ou non limitée, alors que sont isolés l'intérim, l'emploi saisonnier ou d'aide familial... La référence au statut est absente de la catégorisation. Dix ans plus tard, c'est le statut de l'emploi et le passage par les « dispositifs jeunes » qui prévalent. Entre temps, on est passé par deux

Tableau 2
L'évolution des nomenclatures d'état dans les calendriers du Céreq

<table>
<tr><td>Apprentis sortants 78 interrogation 83</td><td>Niveau V sortants 79 interrogation 83</td><td>Inscrits bac 83 interrogation 88</td><td>Niveau V sortants 86 interrogation 89</td><td>Inscrits bac 88 interrogation 90</td><td colspan="2">Niveau V sortants 89 interrogation 94</td></tr>
<tr><td>scolarisé</td><td>scolarisé</td><td>études à temps plein</td><td>études à temps plein</td><td>études à temps plein</td><td colspan="2">études</td></tr>
<tr><td>apprentissage</td><td rowspan="2">salarié (précision : y.c. apprenti, stage pratique, contrat emploi formation)</td><td rowspan="11">emploi</td><td rowspan="11">emploi (précision : y.c. apprenti, CA, CQ, SIVP engagés)</td><td>emploi stable</td><td colspan="2">apprentissage</td></tr>
<tr><td rowspan="10">salarié, stage pratique, contrat emploi formation</td><td rowspan="10">emploi précaire</td><td rowspan="10">emploi</td><td>fonctionnaire</td></tr>
<tr><td>installé</td><td>installé</td><td>CDI</td></tr>
<tr><td>intérim</td><td>intérim</td><td>CDD</td></tr>
<tr><td>aide familial</td><td>aide familial</td><td>SIVP</td></tr>
<tr><td>saisonnier</td><td></td><td>CQ.CA.CO</td></tr>
<tr><td>engagé</td><td>engagé</td><td>CES, TUC, CLO</td></tr>
<tr><td>service national</td><td>service national</td><td>service national</td><td>service national</td><td>service national</td><td colspan="2">service national</td></tr>
<tr><td>chômage</td><td>chômage</td><td>chômage</td><td>chômage</td><td>chômage</td><td colspan="2">chômage</td></tr>
<tr><td>inactif</td><td>inactif</td><td>inactif</td><td>inactif</td><td>inactif</td><td colspan="2">inactif</td></tr>
<tr><td>formation</td><td>formation</td><td>formation</td><td>formation</td><td>formation</td><td colspan="2">formation</td></tr>
</table>

NB : CA contrat d'adaptation, CDD contrat à durée déterminée, CDI contrat à durée indéterminée, CES contrat emploi-solidarité, CLO contrat local d'orientation, CO contrat d'orientation, CQ contrat de qualification, SIVP stage d'initiation à la vie professionnelle, TUC travaux d'utilité collective.

versions minimalistes du calendrier (indication de périodes d'emploi sans autre précision, ou avec la seule distinction entre emploi stable ou précaire).

On peut bien sûr imaginer d'autres catégories d'occupation, construites sur le temps de travail, le salaire, la qualification, ou intégrant les situations multiples. On voit bien qu'alors on aboutirait à de tout autres trajectoires d'insertion. Ainsi, la grande majorité des travaux se glisse dans une nomenclature prédéfinie quels que soient leurs présupposés théoriques.

La nomenclature des caractéristiques individuelles

Les caractéristiques individuelles interviennent comme variables explicatives ou à tout le moins illustratives de la « qualité de l'insertion » (prise comme variable de résultat dans les modèles). Selon la nomenclature qu'on aura choisie pour repérer ces caractéristiques individuelles, on fera apparaître diversement leurs liens avec la variable de résultat. Par exemple la spécialité de formation au premier niveau de diplôme professionnel (niveau V), selon qu'elle est résumée à la seule indication du domaine (industriel ou tertiaire), ou précisée plus finement, n'aura pas le même « effet ». Lorsqu'à diplôme égal, les perspectives d'emploi à l'issue de deux spécialités du même domaine peuvent être sans commune mesure (comme par exemple avec un diplôme d'électronique ou de couture), on ne peut plus postuler l'homogénéité interne d'une catégorie qui regroupe ces spécialités. De plus, une même rubrique de la nomenclature de spécialités peut recouvrir des contenus de formation différents selon la filière : il en est ainsi des spécialités de la santé qui, en apprentissage, désignent des formations au métier de préparatrice en pharmacie, alors qu'en lycée professionnel il s'agit de formations au métier d'aide-soignante... Enfin lorsqu'on sait que certaines spécialités sont préparées quasi-exclusivement en CAP, d'autres en BEP, certaines par l'apprentissage et d'autres par la voie scolaire, on voit l'importance à accorder au contenu des nomenclatures, et aussi aux interactions « cachées » entre les variables. Ainsi les liaisons non linéaires entre diplôme, filière et spécialité invalident-elles l'hypothèse d'additivité : dans un tel cas, et comme souvent, il semble préférable de construire une variable croisée.

Toute nomenclature est une grille réductrice, simplificatrice et critiquable : mais il n'y a pas d'analyse quantitative sans nomenclature. Il revient à l'expert de déterminer la grille la plus appropriée compte tenu de son objet d'analyse, et d'en contrôler les implications sur les résultats au moment de l'interprétation.

DES CODAGES OU LA MISE EN FORME DES CALENDRIERS INDIVIDUELS

La plupart du temps, les variables envisagées pour le traitement ne sont pas directement accessibles dans les calendriers individuels. Elles sont alors construites et cette remise en forme des colonnes du calendrier peut s'entendre comme un recodage préalable au calcul statistique.

Lorsque l'état actuel des calendriers individuels est jugé suffisant pour la suite de l'analyse, la donnée brute est prise telle quelle, sans transformation, nous parlerons alors de *codage neutre* [10]. Une des utilisations des calendriers individuels « à l'état brut » est statique et consiste à extraire l'état d'une cohorte à un moment donné, et à réduire ainsi l'information par des coupes transversales. Ce point de vue permet de comparer les performances de différentes formations à l'aide d'indicateurs globaux (taux de chômage ou d'emploi stable à une certaine date).

Un autre type de codage, que nous nommerons *codage a priori* [11], consiste à procéder à des

Différents codages des calendriers individuels (exemple pour une année)

La donnée de base : FFFEEEEECCEE

Codage neutre FFFEEEEECCEE

Exemple de *codage a priori* : durée totale en Emploi = 7
durée de Chômage = 2
durée en Formation = 3
...

Autre exemple de *codage a priori* :

F3 E5 C2 E2

Exemple de *codage ajusté* : F court, E long, C court, E court

Autre exemple de codage ajusté :

chômage < 3mois
emploi > 6mois

[10] Voir l'encadré : « Différents codages des calendriers individuels (exemple pour une année) ».

[11] Voir l'encadré : « Différents codages des calendriers individuels (exemple pour une année) ».

opérations aveugles sur les données. Ainsi lorsque l'on compte le nombre de passages par une situation d'intérêt (nombre de périodes de chômage ou d'emplois). De tels indicateurs sont encore insuffisants pour rendre compte du processus de l'insertion dans la mesure où ils font abstraction aussi bien de l'ordre que des durées qui caractérisent les calendriers individuels. On peut alors compléter ce résumé descriptif en calculant des durées, variables « sommes » de périodes successives passées dans une situation d'intérêt particulière : durée de chômage, temps d'accès à l'emploi ou durée en formation... Dans certains cas les variables recodées résultent d'une sélection pertinente du calendrier d'états, dans d'autres cas elles donnent une image de la totalité du calendrier. Certaines analyses typologiques procèdent à une périodisation du calendrier d'états qui multiplie les points de repère et permet de réintroduire la dynamique temporelle. Ces opérations ont un caractère automatique et systématique. En tout état de cause, le nombre de périodes retenues et leur longueur modifient peu les résultats des classements ultérieurs.

À la différence des précédents, les autres types de recodage sont dépourvus de ce caractère de reproductibilité complète. En effet, ils relèvent d'un traitement empirique plus collé aux données et qui se donne pour but d'en exprimer les particularités. On peut les nommer *codages a posteriori* ou *codages ajustés* aux données [12]. Ils traduisent les calendriers en termes de successions de combinaisons d'états (emploi, inactivité...) et de durées (« courtes » ou « longues ») en conservant, là encore, tous les épisodes retracés par les calendriers. La validité ne repose que sur l'appréciation de leur pertinence fondée sur l'expertise exogène. En ce sens, même si les procédures sont bien décrites et permettent à un autre expérimentateur d'aboutir à un résultat identique sur l'échantillon concerné, la transférabilité du protocole à une autre cohorte dépend de la stabilité des conditions extérieures.

Toutes les formes de codage butent sur des limites liées au questionnement. En effet, le calendrier d'états tel qu'il est défini *a priori* est limité dans le temps ; la fin de la période de l'insertion professionnelle est conceptuellement mal définie, et son observation est floue. Aussi bien les variables simples que les variables composites peuvent être mal mesurées. Dans le cas des variables simples, cette question des données censurées est l'une des justifications de l'utilisation des modèles de durée (voir *infra*). Pour les autres types de variables, formellement, le problème n'est pas traité.

La mise en forme des données des calendriers individuels conduit donc à définir des variables ; leur rôle dans la suite sera aussi bien celui de variables expliquées, qu'intermédiaires, ou même explicatives.

LES NIVEAUX MULTIPLES DE L'INTERVENTION DU TEMPS

Jusqu'à présent, nous n'avons pris en compte qu'à la marge le fait que les calendriers individuels traduisent l'évolution temporelle des situations individuelles. Le temps est une dimension particulière. C'est souvent par le biais des processus temporels que des éléments de causalité sont introduits dans l'analyse [13]. Il faut donc, là aussi, essayer de mettre à plat les différents traitements. Ceci repose, bien entendu, sur la conception sous-jacente à tous ces travaux, que l'insertion professionnelle est un processus.

Sont alors considérées :

- soit la date à laquelle intervient un événement, par exemple l'entrée dans l'emploi ou dans le chômage, c'est-à-dire aussi un changement d'état, une *transition*. Les dates sont définies à partir d'un repère qui, pour l'insertion des jeunes, peut être marqué par la date calendaire de la sortie du système éducatif [14] ;

- soit la *durée* qui sépare deux événements, par exemple la durée d'accès à l'emploi, temps compté entre la sortie du système éducatif et le premier emploi, ou la durée de séjour, temps qui sépare la sortie de l'entrée dans un état.

Le temps peut servir à construire des *variables de contexte* (« explicatives »). Les caractéristiques des individus – sexe, niveau de diplôme – peuvent être complétées par des indicateurs qui intègrent des durées – avoir eu plus de six mois de chômage au total ou dans la séquence précédant la transition – ou des événements (*i.e.* c'est la « troisième transition »...) Il s'agit alors d'un temps passé que l'on considère comme une variable certaine.

Le temps est surtout *mesure* d'une variable d'intérêt (ou variable « à expliquer »). Ce peut être une variable simple, telle la durée d'accès à l'emploi ou

[12] Voir l'encadré : « Différents codages des calendriers individuels (exemple pour une année) ».

[13] Voir l'ouvrage de Courgeau et Lelièvre, 1989, ou l'article de Lelièvre, 1992.

[14] A la difficulté près que l'entrée réelle dans le processus n'est mesurée que de façon approchée par cette date.

la durée de chômage : on cherche alors à en estimer les caractéristiques de distribution. Le temps est aussi support non seulement d'une mesure mais d'un ordre, éléments de construction d'une *variable complexe*, en l'occurrence la trajectoire, qui traduit des successions et des durées ordonnées (chômage suivi d'un emploi stable,...).

La probabilité conditionnelle qu'un événement se produise à un moment donné sachant qu'il ne s'est pas encore produit, (le hasard ou le risque selon les auteurs), et la loi qu'on lui attribue, conditionnent toutes les estimations qui peuvent être faites par la suite. Là, le temps en est *la source*. Ou bien on suppose simplement qu'il y a une distribution de probabilité non paramétrique dont les observations empiriques permettront de donner une représentation. Ou bien on formule des hypothèses sur ces lois, hasard constant dans le temps par exemple, ou log-linéaire, d'où l'on déduit les lois des variables de durée (exponentielle ou log-logistique...) Dans cette fonction de hasard qui représente en un sens le rôle de la durée, on intègre tous les éléments de causalité qu'on ne maîtrise pas.

L'APPORT DE LA STATISTIQUE

L'ÉVALUATION, UN CADRE DE RÉFÉRENCE

Les objectifs des études analysées se déclinent essentiellement selon le registre de la description explicative ; ainsi peut-on relever les formulations suivantes : *trouver des types* de trajectoires, *reconstituer les parcours*, *expliquer* les différentes trajectoires, *mettre en évidence les facteurs qui accélèrent* (ou retardent) l'entrée dans le premier emploi, *estimer les effets* du passage par un emploi atypique sur la durée d'accès au CDI, du passage par le chômage sur la probabilité de le connaître à nouveau. Quel que soit le cadre théorique auquel ces travaux se réfèrent, leur articulation avec le champ de l'insertion est encore très lâche et les études participent plutôt à la construction de cette articulation qu'à sa mise à l'épreuve.

Le cadre évaluatif que nous avons repéré comme trame commune à l'ensemble des études du corpus nous permet de donner une représentation unifiée des démarches analysées, synthétisée dans un « schéma général de l'évaluation de l'insertion » (tableau 2). On y distingue trois étapes : l'identification du critère d'évaluation (la variable de résultat) ; l'analyse de ses liens avec les caractéristiques « explicatives » ; enfin l'énoncé des conclusions dans la phase d'interprétation. Les encadrés « Un exemple d'application des modèles de durée » et « Un exemple de chaîne typologique » résument des exemples extraits d'études qui ont servi de base à cette réflexion.

La variable de résultat

Dans la démarche économétrique, elle est identifiée, qu'elle soit construite ou disponible dans les données. Il s'agit, par exemple, de la durée de chômage ou de la durée d'accès à l'emploi, de la probabilité d'être au chômage..., variables approchant les dimensions de l'insertion retenues dans la construction théorique. L'identification de cette variable de résultat est un préambule aux calculs.

La recherche de la variable de résultat constitue *le cœur de la démarche typologique*. Cette variable synthétise la trajectoire, qui est alors traduite par l'appartenance à une classe ou la position sur un axe factoriel [15].

L'analyse des liens avec les variables « explicatives »

Cette étape constitue *le cœur de la démarche économétrique* ; l'objet du modèle est de quantifier les effets des variables explicatives ou tout au moins d'en tester l'existence et le sens. On suppose que le modèle est complet et exhaustif, l'amplitude de l'effet d'une variable est conditionnelle à la présence des autres variables.

Dans l'analyse typologique, l'établissement de ces liens passe la plupart du temps par l'examen des croisements de la variable de résultat avec les

Tableau 3
Schéma général de l'évaluation de l'insertion

	Modèle économétrique	Modèle typologique
Variable de résultat	identifiée : cette étape est un préambule	recherchée : *cette étape constitue le cœur de la démarche*
Analyse des liens avec les « variables explicatives »	quantification des effets *cette étape constitue le cœur de la démarche*	- écarts à l'indépendance - régression logistique
Interprétation	- détermination explicite de la causalité - en termes de probabilité : interférence à partir de l'échantillon	pas de référence explicite à la représentativité de l'échantillon

[15] Bien que n'ayant pas rencontré dans les études de cas explicite d'utilisation des facteurs comme variable synthétique, nous le mentionnons ici : ce point sera développé plus loin dans la discussion de la chaîne typologique.

variables explicatives. Une étape supplémentaire est parfois franchie avec la quantification des effets par une régression logistique sur la variable de résultat.

L'interprétation

Dans la première démarche, elle conduit souvent à la détermination explicite de causalité d'autant qu'il y a un effet temporel. On infère les résultats à la population parente à partir de l'échantillon. Ils sont exprimés en termes de probabilités [16].

En revanche, ils sont moins souvent exprimés en termes de causalité dans le deuxième cas mais les écarts à l'indépendance traduisent aussi des effets des variables explicatives, plus difficiles à quantifier.

On voit que les deux démarches diffèrent radicalement quant à leur objet, et la place qu'elles occupent dans le schéma d'ensemble : la préoccupation principale, dans la démarche typologique, est de chercher des régularités dans la diversité des trajectoires, pour en proposer une synthèse qui aura statut de variable de résultat. C'est donc une étape intermédiaire entre la mise en forme des données et l'estimation d'un modèle : mise en forme de la trajectoire qui aboutit en fait, comme on le verra plus loin, à proposer une variable de résultat synthétique élaborée. Ce qu'on modélise, c'est la trajectoire, et non ses liens avec d'autres variables individuelles. Alors que dans la démarche économétrique, la variable de résultat est identifiée d'emblée, le critère de l'évaluation est clairement posé, et la modélisation porte sur ses liens, plus ou moins sophistiqués, avec d'autres caractéristiques de l'individu.

Certains vont utiliser une partition de l'échantillon obtenue par classification comme une nouvelle variable discrète, dont ils cherchent les déterminants par la méthode de la régression logistique. C'est un exemple d'enchaînement des deux démarches.

LES TYPOLOGIES DE TRAJECTOIRES

Les typologies de trajectoires ont connu depuis plus de dix ans dans le domaine de l'insertion un succès croissant. Elles reposent sur des classifications qui visent à rechercher un ordre dans une réalité complexe.

[16] Dans l'énoncé des conclusions, il y a parfois un glissement sémantique tel qu'on ne sait plus au juste quel est l'espace probabilisé : le résultat « on a une probabilité p d'observer l'évènement (voire le comportement) X parmi les individus ayant la caractéristique Y » devient « un individu ayant la caractéristique Y *a la probabilité* p de connaître l'événement (*ou même d'avoir le comportement*) X ». On est passé des événements concernant la population d'individus, aux événements du futur de l'individu...

Par exemple, avec un calendrier de sept états et de 40 mois, il y a potentiellement 7^{40} calendriers individuels possibles et donc 7^{40} trajectoires. En réalité les trajectoires ne sont évidemment pas produites par le hasard, et la diversité observée est beaucoup moins foisonnante, d'autant qu'on l'aura le plus souvent réduite par recodage (voir plus haut la mise en forme des trajectoires). Elle est cependant encore trop riche pour se laisser appréhender sans recourir à des simplifications : la classification va permettre, en assimilant des trajectoires très semblables *au sens de la distance choisie*, des mises en équivalence qui faciliteront la mise en relation entre les faits sociaux qu'on cherche à décrire et même à expliquer (les trajectoires d'insertion) et leurs déterminants.

Il n'y a pas *une* typologie de référence. Cela n'est pas gênant en soi, pas plus que d'avoir plusieurs mesures d'un même objet : il faut seulement expliciter les conditions de la mesure. En effet, le résultat d'une classification dépend de la distance et de la méthode d'agrégation choisies. Ainsi, dans les exemples rencontrés, on procède à une classification hiérarchique, suivie dans certains cas d'une agrégation autour de centres mobiles pour stabiliser les classes. Cette classification est élaborée directement à partir des descripteurs de la trajectoire ou à partir des résultats d'une analyse factorielle sur ces trajectoires. Le critère d'agrégation est le critère classique de Ward (maximisation à chaque pas de l'inertie interclasse). La distance est une distance euclidienne opérant sur des tableaux plus ou moins proches du tableau des calendriers individuels.

Si la démarche classificatoire vise d'abord à décrire et résumer une réalité complexe, elle est aussi construction d'une nouvelle variable qui sera simplement croisée avec les caractéristiques individuelles des classes, ou entrée dans un modèle de régression logistique pour expliquer l'appartenance aux classes.

Cette construction est en elle-même féconde en ce qu'elle force à interpréter les regroupements et leur mode d'obtention. Les étapes n'en sont cependant pas encore suffisamment explicitées et reproductibles pour que le résultat soit réellement opératoire et accède au statut de nomenclature robuste.

La question de la stabilité de la partition tient essentiellement à deux aspects : sa robustesse et sa reproductibilité. La *robustesse* des typologies est confirmée si les résultats restent stables dans des épreuves de simulation aléatoire. On peut s'assurer, par exemple, de l'exhaustivité du tableau soumis à l'analyse par suppression de lignes ou de colonnes, de la pertinence

Un exemple de chaîne typologique

Canevas extrait de A. Degenne, M.-O. Lebeaux et L. Mounier, *Construction d'une typologie de trajectoires à partir de l'enquête de suivi des jeunes des niveaux V, Vbis et VI* (Caen 1995).

L'objectif des auteurs est de repérer les grandes tendances de la mise en ordre des itinéraires individuels. Il se dit purement exploratoire et descriptif mais les auteurs soulignent que « *comme dans toute analyse des données, le mode de description choisi introduit des hypothèses implicites ou explicites qui pèsent sur les résultats* ». Ainsi en est-il du rôle des transitions (élément de la stratégie des acteurs) et de celui des réseaux.

Mise en forme des données

La nomenclature de base compte neuf états qui sont ceux du calendrier enrichis du statut de l'emploi. Le calendrier de 42 mois est découpé en sept séquences de six mois. Pour chaque séquence on compte le nombre de mois passés dans chaque situation, soit 7x9 = 63 variables. On compte aussi le nombre d'occurrences de chacune des transitions d'un état à un autre, soit 9x9 = 81 variables. On a donc en tout 144 variables décrivant chaque individu.

Traitement statistique

Analyse factorielle du tableau, puis classification ascendante hiérarchique fondée sur les 50 premiers facteurs. Les classes sont ensuite stabilisées par réaffectation des individus par la technique des centres mobiles. On retient cinq groupes d'itinéraires.

Interprétation

L'interprétation des groupes se fait par l'étude des trajectoires moyennes. Ensuite les groupes sont expliqués par des variables exogènes. Des effets, rôle du diplôme, de la situation des parents,... sont reconnus.

du champ des variables retenues par perturbation des colonnes du tableau, et de la force de liaison entre variables explicatives et variables expliquées par réaffectation aléatoire des valeurs des variables expliquées mais ces techniques ne sont certainement pas utilisées de façon systématique par les chercheurs et on s'en remet à une validation souvent partielle de la procédure de calcul.

Le problème de la *reproductibilité* est, lui, commun à toute la statistique inférentielle. En effet, dans la démarche typologique on cherche une structure qui serait présente dans la population et dont l'échantillon, en tant qu'observation empirique de cette réalité, permet de donner une image. On se trouve dans un cas d'application de la loi des grands nombres, dont les hypothèses sont suffisamment peu contraignantes pour qu'elle soit utilisée ici. On a une situation équivalente à celle de l'ajustement d'un modèle à la réalité. Il n'est pas plus trivial de dire que le modèle est découvert en même temps que mis à l'épreuve, que de dire que la valeur la plus probable de la moyenne est celle trouvée dans l'échantillon. L'échantillon étant supposé représentatif, la variable de résultat n'a pas de raison d'être biaisée et la typologie doit être reproductible sur un autre échantillon de la même population.

En tout état de cause, on n'utilise la partition comme variable de résultat que lorsque la séparabilité des types est assurée. Cette séparabilité est estimée sur l'échantillon. Deux cas extrêmes balisent la réalité : celui de la séparabilité absolue, où la variance intraclasse est seulement une variance d'échantillonnage, et celui où les frontières entre les types sont floues et les centres de classe ne sont que des repères dans un espace de dimension réduite (l'espace des facteurs). Dans un cas on aura une variable discrète, dans l'autre un petit nombre de variables continues.

Dans ce dernier cas, si on a pu identifier un facteur comme facteur de qualité de l'insertion, on a alors une variable de résultat quantitative qui peut entrer dans un modèle économétrique. Dans le cas où la variable de résultat est discrète, elle n'est cependant pas ordinale, ce qui complique l'évaluation. On perçoit pourtant, derrière les commentaires qui en sont donnés, une hiérarchisation implicite des types de trajectoires, calquée sur une hiérarchisation – implicite elle aussi – des statuts d'occupation à partir desquels sont décrits les parcours :

inactivité < chômage < mesure jeune < CDD < CDI.

Remarquons enfin que dans la plupart des études, dont l'objectif principal est l'identification de la

variable de résultat, la démarche typologique est rarement achevée par le recours aux techniques de l'analyse des données pour l'étude des liens entre variables (analyse du tableau de régression, utilisation de la technique des éléments supplémentaires...)

LA MODÉLISATION ÉCONOMÉTRIQUE

On peut noter dans ce cadre de l'évaluation de l'insertion à partir d'enquêtes longitudinales que la démarche économétrique est la plupart du temps essentiellement statistique. Ainsi aux décisions qui concernent les nomenclatures s'ajoute un certain nombre de décisions qui permettent d'adapter le meilleur modèle aux données observées.

Sans entrer dans le détail des méthodes utilisées, on peut relever les différentes options qui balisent le processus d'estimation. Elles concernent le choix du modèle, celui des variables explicatives, la forme du hasard et le mode d'action des caractéristiques individuelles. Ainsi, par exemple, dans le cas de l'estimation des durées d'accès à l'emploi, la forme du hasard (constant ou non, monotone ou non) est choisie en fonction des observations ; vient ensuite l'estimation d'un modèle non paramétrique, ou d'un modèle paramétrique avec une loi des durées de type log-linéaire ou exponentiel... La variable est précisée (l'événement est l'entrée en emploi), l'origine du temps est la même pour tous – sinon les durées individuelles sont ramenées, par convention, à une même origine.

Ces procédures doivent être adaptées lorsque les données proviennent d'enquêtes faites par sondage dans un stock (de chômeurs, par exemple), en corrigeant les estimations pour tenir compte du biais d'échantillonnage. Mais dans la plupart des études recensées ici, la question ne se pose pas. Quant à la censure à droite des données, ses difficultés sont résolues par l'utilisation des modèles de durée.

Le rôle des caractéristiques individuelles observées est estimé lui aussi par leur introduction dans le modèle de base sous des formes diverses (modèles à risques proportionnels constants, à vie accélérée...) Dans ce registre, ce sont des considérations liées aux observations qui guident le choix des spécifications les plus efficaces.

Un exemple d'application des modèles de durée

Schéma extrait de J.-M. Le Goff, *Processus d'accès à un emploi sur contrat à durée indéterminée de jeunes sortis de terminale en 1983* (Caen 1995).

L'auteur s'inscrit dans une recherche sur la forme des processus biographiques qui combinent deux types de déterminisme. Le premier se rapporte au rôle joué par la formation initiale en tant qu'événement fondateur : d'une part elle est censée jouer un rôle presqu'indépendant de la personne qu'elle qualifie et d'autre part elle peut s'effacer au cours du temps. Le second type consiste à considérer qu'une situation à un moment donné dépend des étapes antérieures. On se trouve donc dans l'hypothèse d'un schéma de relations de cause à effet.

Mise en forme des données

Calcul des durées d'accès au premier emploi : soit atypique (D1), soit en CDI (D2), en défalquant la durée du service militaire. Il s'agit d'un codage *a priori*.

Traitement statistique

Utilisation d'un modèle à risques proportionnels pour estimer l'impact des différentes caractéristiques individuelles sur les probabilités d'accès direct au CDI comparées à l'accès indirect (avec passage par un emploi atypique considéré comme l'événement perturbateur).

Interprétation

Le passage par un ou plusieurs emplois atypiques retarde l'accès à un CDI.

La série du bac d'origine joue un rôle sur les chances d'accès direct au CDI pour ceux qui entrent directement dans la vie active et non pour ceux qui tentent des études supérieures.

L'effacement du rôle fondateur de la série de terminale sur les chances d'accès au CDI est accompagné par l'émergence du rôle du sexe.

Si les emplois sont distingués par la nature du contrat de travail, ce sont les faits d'itinéraire les plus récents qui sont pris en compte par les employeurs et non la formation suivie en terminale.

Les méthodes sont naturellement plus complexes lorsqu'on considère les transitions entre plusieurs états ou la récurrence des passages par un état donné. La probabilité d'une trajectoire individuelle peut être formalisée. Mais les modèles qui rendraient compte de l'hétérogénéité individuelle, en même temps que de l'ensemble de la dynamique temporelle, sont si complexes et difficiles à estimer sur les échantillons dont on dispose, qu'il faut en réduire la dimension en assimilant des paramètres ou des effets, en simplifiant la dépendance temporelle. Cela conduit à utiliser des modèles markoviens ou semi-markoviens.

En fin de compte, la plupart des hypothèses sont justifiées par la recherche d'une bonne adaptation aux données ou la faisabilité de l'estimation. On voit là aussi le rôle de l'ajustement qui a été mis en évidence dans le traitement statistique des typologies.

* *
*

Le cadre évaluatif commun à l'ensemble de ces travaux implique une prise en compte de la *causalité* : on s'attend à ce que les résultats soient exprimés en termes d'effets et dans la mesure où le temps est le support de ce type de données, on peut introduire naturellement une forme de relation causale temporelle. Cette attente ferait plutôt pencher la balance du côté d'une analyse « explicative » mais l'opposition entre méthodes explicatives et descriptives existe-t-elle réellement ?

Le terme d'explicatif est souvent utilisé de façon abusive, et celui de descriptif est péjoratif. Il faudrait comprendre que « descriptif » correspond à une démarche dans laquelle l'objectif est de *re-présenter* les données : les « présenter » autrement, pour en assurer une meilleure lisibilité. Histogrammes, axes factoriels, classes, sont plus lisibles que le tableau brut ; ils contribuent à construire de nouvelles variables. Alors que dans les modèles de type économétrique, une partie des données Y est à *reconstituer* à partir d'autre(s) partie(s) X. Le vocable « reconstitution de données » paraît plus neutre qu'« explicatif » et correspond mieux à la véritable nature scientifique de ces méthodes statistiques.

La modélisation économétrique aussi bien que l'analyse typologique participent de ces deux démarches : que l'on pense seulement à l'utilisation des éléments supplémentaires comme outils décisionnels ou au rôle des méthodes non-paramétriques dans les modèles. La frontière *technique* entre les deux pôles « explicatif » et « descriptif », tombe ainsi souvent d'elle-même.

La sûreté de *l'interprétation*, et donc des conséquences en termes d'évaluation que l'on en tire, dépend du niveau de *validité* qui peut lui être attribué. Cette validité peut être interne ou externe. Dans les modèles, la validité est généralement traitée à l'intérieur du modèle lui-même : les modèles examinés ici ont chacun leurs exigences de validité (en termes d'hétérogénéité, de censure à droite, à gauche, de loi du hasard, de termes normaux...) Pour répondre aux impératifs de l'action, on devrait aussi y adjoindre des procédures de validation externe. Quant aux typologies, leur validité reste qualitative, c'est-à-dire interne à la chaîne de travail. La partie quantifiable, externe aux procédures, fonctionne essentiellement par simulations et est la plupart du temps omise. Il y a donc une sous-exploitation de ce type de démarche qui devrait être compensée pour que des variables complexes soient utilisables dans un schéma prédictif. En tout état de cause, à côté de ces aspects statistiques, l'interprétation repose toujours sur des éléments de connaissance extérieurs aux procédures statistiques.

Ainsi l'articulation des méthodes nous semble-t-elle ouvrir une voie pour dépasser les difficultés d'analyse soulevées par la complexification croissante des processus d'insertion et pour en assurer la compréhension. Mais le travail préliminaire au traitement statistique constitue, comme on le sait bien, une phase importante dont l'impact est fondamental. Pour répondre aux besoins de clarification évoqués au début de cet article, il est nécessaire que les rapports d'étude explicitent très rigoureusement *l'ensemble des conditions et des opérations* qui ont abouti à la production des résultats. On peut aussi imaginer, pour préciser l'impact des orientations prises à chacune des différentes étapes repérées, de monter une expérimentation sur un corpus de données unique.

Jean-Pierre Fénelon
LES
Yvette Grelet
Céreq
Yvette Houzel
LES

Bibliographie

Actes des trois premières journées du longitudinal, Toulouse, Document Céreq, n° 99, septembre 1994 (eds. M. Ourtau et P. Werquin) ; Caen, Document Céreq, n° 112, décembre 1995 (eds. A. Degenne, M. Mansuy et P. Werquin) ; Rennes, Document Céreq, n° 115, juillet 1996 (eds. A. Degenne, M. Mansuy, G. Podevin et P. Werquin).

Aucouturier A.-L. (1994), *Panels et évaluation des politiques d'emploi, Cahiers Travail et Emploi*, La Documentation française, Paris.

Courgeau D., Lelièvre E. (1989), *Analyse démographique des biographies*, INED.

Desrosières A. (1996), *Les apports mutuels de la méthodologie statistique et de la sociologie*, Conférence inaugurale des V[e] Journées de Méthodologie statistique, INSEE.

Elbaum M. (1997), *Allocution d'ouverture aux quatrièmes journées du longitudinal*, Document Céreq n° 128, octobre.

Fargher S. (1997), *The Economic Effect of Greater Diversity in Training*, Communication aux journées du Network on Transitions in Youth, Dublin.

Gautié J. (1996), *L'évaluation de la politique de l'emploi en faveur des jeunes en France*, Dossier du CEE n° 8.

Lechêne V. et Magnac T. (1996), « L'évaluation des politiques publiques d'insertion des jeunes sur le marché du travail », in *Les jeunes et l'emploi, Cahiers Travail et Emploi*, La Documentation française, Paris.

Lelièvre E. (1992), « L'étude des interactions entre phénomènes : dépendance unilatérale et causalité », in *Démographie et différences*, PUF coll. AIDELF.

Schömann K. (1996), *Longitudinal designs in Evaluation Studies*, in *Handbook of Labour Market Policy and Policy Evaluation*, Berlin, Wissenschaftzentrum.