Net.Doc . 82

Les parcours d'insertion des jeunes : une analyse longitudinale basée sur les cartes de Kohonen

Patrick Rousset
Céreq
rousset@cereq.fr

Jean-François Giret
Institut de recherche sur l'éducation : sociologie et économie de l'éducation (IREDU)
Centre associé régional du Céreq - Bourgogne
jean-françois.giret@u-bourgogne.fr

Yvette Grelet
Espace et Sociétés (ESO), Université de Caen
Centre associé régional du Céreq - Basse-Normandie
yvette.grelet@unicaen.fr

Céreq,
10, place de la Joliette, BP 21321,
13567 Marseille cedex 02.

Avril 2011

SYNTHÈSE

Cet article présente une méthode pour élaborer des typologies d'itinéraires individuels qui prenne en compte la dynamique longitudinale des parcours. Elle s'applique lorsque ces itinéraires sont décrits à partir de calendriers. L'originalité de cette méthode réside à la fois dans le mode de calcul de la distance entre trajectoires, et dans la procédure de classification basée sur les cartes d'auto-organisation de Kohonen. La propriété principale de la distance est de prendre en compte de la proximité entre les états et son évolution dans le temps.

Cette méthode est appliquée ici à l'analyse des trajectoires professionnelles, à partir de données du Céreq permettant le suivi mensuel sur sept ans de jeunes sortis en 1998 du système éducatif. Les résultats mettent en évidence l'importance de la dynamique temporelle dans la construction des parcours d'entrée dans la vie active. La place et les apports de cette méthode sont enfin comparés aux plus couramment employées pour construire des typologies de parcours.

SOMMAIRE

INTRODUCTION

Cet article s'inscrit dans le cadre de l'analyse des données longitudinales. Il propose une méthode de classification des trajectoires de vie et son application à une typologie des parcours d'insertion sur le marché du travail. Les méthodes typologiques sont fréquemment utilisées en compléments ou alternatives aux modélisations économétriques, pour analyser les parcours individuels. Elles s'inscrivent dans une approche holistique qui met l'individu au centre de l'analyse, dans sa globalité et dans sa dynamique. Les typologies permettent de résumer une grande variété de trajectoires en un petit nombre de types et de relier ces types de parcours avec les caractéristiques dominantes des individus qui les suivent.

Le récit de vie d'un individu est généralement une suite dans le temps d'évènements datés, de durées différentes. Ces évènements ont comme caractéristiques d'être homogènes ou stables dans le temps et s'interrompent par une transition vers un autre évènement. Dans le cadre de l'insertion sur le marché du travail, un évènement est une expérience en emploi ou hors emploi. Pour analyser ces récits, les données biographiques sont recodées et nous nous intéresserons dans la suite au format de recodage sous la forme d'un calendrier, même si la méthode présentée ici peut s'adapter à d'autres structures. Comme toute forme de codage, le calendrier relève de choix arbitraires : d'un découpage identique pour tous les individus de la période d'observation et d'une recomposition des positions des individus en un ensemble fini d'états possibles. C'est donc une transformation de l'information qui rompt dans sa forme avec le mode récit décrit plus haut. Dans l'exemple que nous traiterons ici, le calendrier est un relevé mensuel sur 7 ans de l'état sur le marché du travail de 16000 jeunes.

L'arbitraire de ces choix de recodage n'est pas anodin et pourrait influencer le résultat. La structure temporelle du récit initial en termes de stabilité et transitions est perdue par le mode de découpage du temps. L'ordre séquentiel des unités de temps (les fenêtres) peut permettre d'en reconstituer une autre mais il appartient à la méthode d'analyse de le faire. De même, dans la pratique, le choix de l'ensemble fini des états résulte souvent d'un regroupement arbitraire. Celui-ci est directement lié au sujet de l'étude. Ainsi, dans le cadre de l'insertion professionnelle, il est traditionnel de regrouper le contrat à durée indéterminé avec le statut de fonctionnaire alors que dans une analyse des statuts sur le plan législatif ils auraient été distingués. Ici aussi, il convient à la méthode de tenir compte de ces arbitrages pour des raisons de cohérence et pour éviter que le résultat résulte directement d'un effet indésirable de cet arbitraire.

L'originalité de notre méthode est de reconstituer la dynamique des récits. Appliquée à l'étude de l'insertion, la méthode révèle une structure du marché de l'emploi à partir des transitions entre les états pour chaque date et chaque pas. Elle tient de plus compte de l'échéance de la transition de sorte qu'une transition à court terme compte d'avantage qu'une transition à long terme. Parmi les propriétés attendues figure la capacité à prendre en compte la structure spécifique à la période d'insertion -les transitions d'emplois temporaires vers emploi durables- ainsi que celle liée à la conjoncture -périodes de crise et de croissance-. Par exemple pour le contrat à durée déterminée, son rôle intégrateur vers le contrat à durée indéterminée évolue au cours des 7 années post-études, alors que sa précarité augmente en période de crise. On mettra ainsi en évidence trois phases d'évolution du CDD et deux de l'intérim, qui voient ces types de contrat aller d'un rôle d'insertion à un rôle de maintien sur les marges du marché du travail. On peut également citer l'impact des périodes de crise et de croissance sur la précarité d'un emploi en CDD.

La méthode que nous développons ici se déroule en trois étapes. La première - la distance entre états et sa dynamique dans le temps - met en évidence la structure du marché du travail des jeunes et son évolution dans le temps, en s'appuyant non seulement sur l'ordre séquentiel de deux évènements mais aussi sur la durée qui les sépare. Elle amènera par exemple à considérer le CDD comme plus proche du chômage dans la période où domine son caractère précaire, qu'il ne l'est lorsqu'il joue plutôt un rôle d'insertion vers le CDI. Dans une seconde étape, la méthode intègre ces résultats pour construire la distance entre les trajectoires. La troisième étape consiste à utiliser les cartes d'auto-organisation pour construire une typologie de trajectoires.

Parmi les autres méthodes, certaines partent aussi d'une structure de l'espace des états mais souvent statique (l'optimal matching) ou qui ne tient pas compte de l'ordre séquentiel (ACM).

La méthode présentée restitue les dynamiques de l'insertion. C'est ainsi par exemple qu'en rapprochant les statuts précaires en amont de la classification, elle permet de faire émerger des trajectoires « descendantes » du CDI vers ces statuts précaires (CDD, intérim ou chômage en fin de période) qui n'apparaissent pas avec d'autres méthodes.

Pour calculer une distance entre trajectoires, il faut d'abord se doter d'une distance entre états, laquelle n'est pas forcément constante dans le temps. Ainsi la distance entre trajectoires que nous proposons répond à l'objectif de prendre en compte l'évolution temporelle de la proximité entre les états. La proximité entre états à l'instant t est mesurée à partir des transitions entre états observées pour chaque pas (entre l'instant t et l'instant t+1, t+2, …t+n) : deux états sont considérés proches si les échanges entre eux sont importants (un grand nombre d'individus passent de l'un à l'autre) ou s'ils comptent les même fréquences de transition vers les autres états. Cette mesure de proximité ne nécessite pas d'information exogène, elle se fait à partir de la seule information contenue dans les trajectoires elles-mêmes. La distance entre trajectoires, reposant sur les transitions, est adaptée au cadre euclidien. Prenant pour critère la proximité entre états et son évolution, nous la comparerons avec les distances couramment utilisées pour l'analyse des calendriers individuels : la distance du χ^2 sur le calendrier (Grelet, 2002; Escofier-Cordier, 2003) ou sur le calendrier agrégé par périodes de temps pour l'analyse harmonique, et celle de l'optimal matching (Abbott et Hrycak, 1990). La première, classique en analyse des données qualitatives nominales, nous servira de référence. La seconde en est une variante. La troisième, importée de la génétique, a été adaptée en sociologie pour mesurer les distances entre séquences d'histoires de vie. Néanmoins il existe un éventail très large d'autres propositions de distances entre trajectoires, dont certaines développées dans le cadre d'analyses des calendriers professionnels des enquêtes du Céreq (Fénelon et alii, 1997).

Pour ce qui est de la méthode de classification, nous en examinerons plusieurs appartenant aux deux grandes familles des classifications hiérarchiques (partitions emboîtées en arborescence), ou des partitionnements par agrégation autour de centres mobiles. Nous montrerons l'apport des cartes d'auto-organisation (Kohonen, 2001) et leur complémentarité avec la distance que nous proposons, l'essentiel étant que leur système de représentation permet de travailler à un niveau très fin, et de faire ainsi apparaître à la fois les proximités entre états et leurs évolutions dans le temps.

L'étude de l'insertion professionnelle de sortants du système éducatif, sur des données issues des enquêtes Génération 98 du Céreq, nous servira d'illustration.

1. LA STRUCTURE DES DONNÉES : LES CALENDRIERS

1.1. Les calendriers

D'un point de vue formel, le *calendrier* d'un individu *i* donne la suite des états qu'il occupe à chaque instant *t*. L'ensemble E des états possibles est fini. Les états seront représentés par des nombres entiers (les codes) prenant leurs valeurs dans {1,2,…,*e*,…,*E*}. Les calendriers individuels constituent les lignes d'un tableau (X_i^t) de dimensions *I*x*T,* où *I* est le nombre d'individus et *T* le nombre d'instants (figure 1). Il s'agit là d'un tableau de données sous forme de codage condensé.

Le tableau des codes sera converti, en préalable à tout traitement, en tableau de nombres, suivant le procédé utilisé en analyse des correspondances multiples (Lebart et al, 2006). Du calendrier, on déduit ainsi le tableau classique **disjonctif-complet** (Y_i^s), à *I* lignes et *T***E* colonnes (figure 2). Le *E*-uplet t décrit l'état à l'instant t : sa $e^{ième}$ composante vaut 1 si l'individu est dans l'état *e* à cet instant, et 0 sinon. Ainsi, chaque colonne *s* du tableau correspond à un couple (*t,e*).

Nous appellerons dans la suite ***situation*** e_t l'état *e* à l'instant *t*. *La situation* sera la notion centrale de la méthode, elle permet de distinguer un même état à deux dates différentes et donc de prendre en compte son évolution dans le temps. Par exemple « *CDI en septembre* 1998 », « *CDI en octobre* 1998 » et « *intérim en septembre* 1998» sont trois situations différentes. Dans le tableau disjonctif-complet (figure 2), *chaque colonne correspond à une situation s*= e_t qui est codée 1 si l'individu se trouve à l'instant *t* dans l'état *e*, et 0 sinon.

De la même façon, nous dirons qu'il y eu une transition entre les situations *s* et *s'* lorsqu'un individu a connu à la fois les situations *s*=e_t et *s'*=$e_{t'}$, donc si l'individu qui était en *e* à l'instant *t* s'est retrouvé en *e'* à l'instant *t'*. Le nombre de transitions entre situations est donc le nombre de cooccurrences entre *s* et *s'*.

Remarque :

Les deux tableaux se correspondent : lorsque X_i^t vaut *e*, $Y_i^{s=e_t} = 1$, et 0 sinon.

Figure 1

Tableau (*X*) de données sous la forme de codage condensé

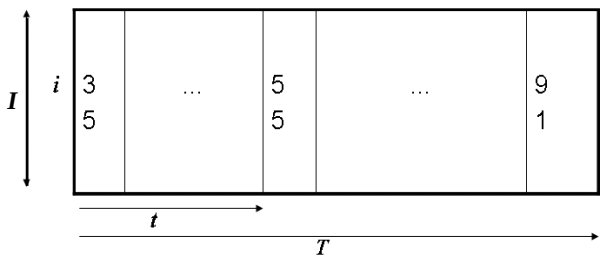
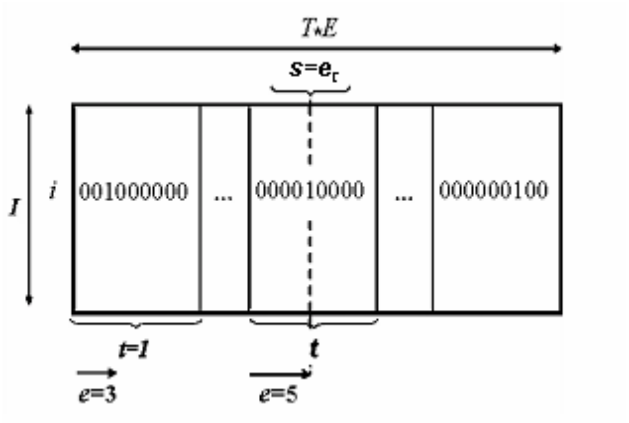


Figure 2

Tableau (Y) de données sous la forme de codage disjonctif-complet

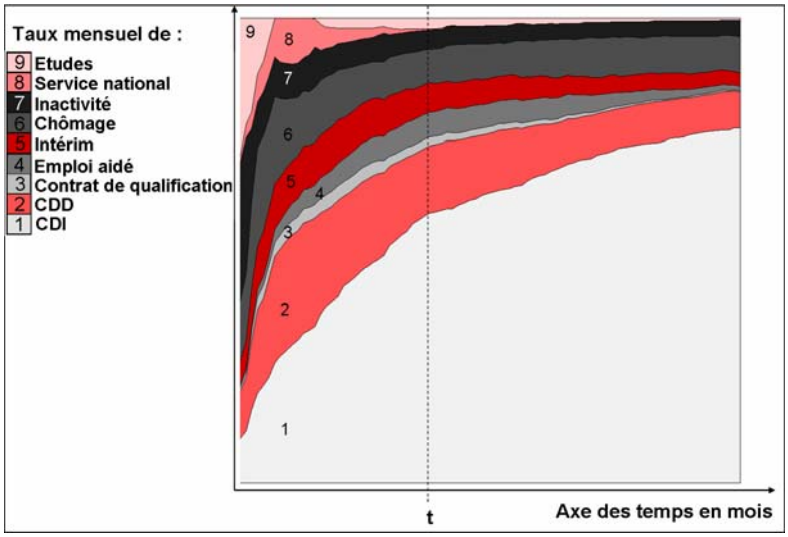


1.2. Un exemple de calendrier : les parcours d'insertion

L'enquête « Génération 98 à 7 ans » du Céreq a interrogé, en 2005, 16 000 jeunes sortis de formation initiale en 1998. Elle reconstitue pour chacun le calendrier mensuel des positions qu'il a occupées sur le marché du travail au cours des 7 années (88 mois), selon une nomenclature d'états en cinq positions d'emploi et quatre positions de non-emploi : *le CDI*, *le CDD*, *l'intérim*, *le contrat de qualification*, *les autres contrats aidés* et respectivement *le chômage*, *l'inactivité*, *le service national* et *la formation ou reprise d'études*. Le graphique suivant (Figure 3), appelé chronogramme, est une juxtaposition des histogrammes cumulés donnant pour chaque mois la part relative de chaque état à cet instant. Il représente ainsi l'évolution dans le temps (axe des abscisses) de la contribution de chaque état au contingent de la cohorte. Ce type de représentation donne une bonne idée de l'évolution dans le temps d'un ensemble d'individus. Par contre, il masque le nombre de transitions individuelles en donnant l'image d'une évolution systématique et régulière.

Figure 3

Chronogramme de l'insertion des jeunes sortis de la formation initiale en 1998 sur 7 ans



1.3. L'analyse des calendriers

En l'absence d'information exogène, la façon de définir les proximités entre états qui paraît la plus adaptée aux données longitudinales est de prendre en compte l'intensité de leurs liens. Par exemple on peut considérer que deux états sont proches si un grand nombre d'individus passent, à des instants différents, par l'un et l'autre de ces deux états. On peut aussi considérer qu'ils sont proches si un grand nombre d'individus qui passent soit par l'un, soit par l'autre passent aussi par un même troisième état. Ainsi si un grand nombre d'individu en *CDD* obtiennent ensuite un *CDI*, ces deux états peuvent être considérés comme proches et si les individus ayant connu un *CDD* comme ceux ayant connu l'*intérim* connaissent par la suite un *CDI*, le *CDD* et *l'interim* seront également proches même si peu d'individus ont connu les deux. A ce stade, on peut faire trois remarques :

- La première est que le nombre d'individus ayant connu à la fois *e* et *e'* dépend des fréquences des états *e* et *e'*. On peut se garantir d'une trop grande sensibilité à la fréquence des états en utilisant une distance qui possède la propriété *d'équivalence distributionnelle*, à savoir que la distance entre lignes (resp. colonnes) est inchangée si on agrège des colonnes (resp. des lignes) de même profil (Lebart *et al.* 2006, Escofier-Cordier 2003).
- La deuxième remarque est qu'on peut affiner la mesure des proximités en introduisant le temps. En effet, que 40 % des individus en *CDD* à t_1 se retrouvent en *CDI* en t_2 n'est pas aussi révélateur de proximité entre *CDD* et *CDI* si l'écart entre t_1 et t_2 est d'un an ou de quatre ans. Dans le premier cas, le lien est plus fort (on parlera d'une incidence directe du passage en *CDD* pour obtenir un *CDI*). Dans le second, les individus peuvent avoir connu nombre de situations intermédiaires.
- La troisième remarque est que les liens entre états évoluent dans le temps. Comme on l'a déjà indiqué, les flux de passage du *CDD* au *CDI* sont plus importants en début qu'en fin d'insertion.

Partant de ces trois remarques, nous chercherons à définir une distance qui respecte le principe d'équivalence distributionnelle et qui mesure la proximité entre les états en incluant l'évolution dans le temps.

2. UNE DISTANCE ENTRE TRAJECTOIRES QUI PREND EN COMPTE L'ÉVOLUTION DANS LE TEMPS

2.1. Structure spatio-temporelle inter-états

Nous cherchons à définir une distance entre trajectoires qui intègre la structure spatio-temporelle de l'espace des états, c'est à dire la structure de l'espace des situations –états indicés par le temps. Dans l'espace des individus, les coordonnées des situations sont les colonnes du tableau disjonctif-complet *(Y)*. Deux situations seront d'autant plus proches que sera grand le nombre d'individus réalisant le code 1 pour chacune d'elles, le nombre de cooccurrences. Cette proximité correspond aux notions classiques de distance entre profils-colonnes dans l'analyse des correspondances multiples. Elle se mesure généralement avec la distance du χ^2. Elle n'inclut pas l'ordre séquentiel comme critère de proximité, et par conséquent, les corrélations à long terme et à court terme sont traitées de la même façon.

$$d^2(s,s') = \sum_{i=1}^{I} n\left(\frac{Y_{is}}{Y_{.s}} - \frac{Y_{is'}}{Y_{.s'}}\right)^2$$

Nous cherchons également à rapprocher deux situations en fonction du nombre d'individus qu'elles partagent. Deux situations seront proches si le nombre d'individus qui connaissent l'une et l'autre est important (par exemple si un grand nombre d'individus connaissent les deux situations « *CDD en octobre 1998* » et « *CDI en octobre 1999* » - cela signe le rôle d'intégration du CDD vers le CDI ; et si les individus qui sont en *CDD ou en intérim en octobre 1998 -deux situations de même date-* se retrouvent en *CDI* en octobre 1999, cela indique le rôle de tremplin pour l'insertion de ces deux types de contrat). Dans le cas de l'insertion, cette approche permet de faire valoir le rôle « intégrateur » de différents types de contrats de travail qui sont, par leurs statuts, transitoires dans la trajectoire.

On peut de plus, prendre en compte l'aspect temporel des liens entre situations en accordant plus d'importance aux liens de court terme que de long terme. Ici, nous considérons en effet que corrélation entre deux situations peu éloignées dans le temps est plus révélatrice d'une incidence ou d'un lien direct que celle à long terme qui dépend des parcours intermédiaires. Il est alors préférable d'utiliser une distance qui distingue court et long terme, par exemple pour évaluer les capacités intégratives des contrats courts.

La distance entre les situations que nous proposons repose sur les transitions (entre une situation et son futur probable) et leur échéancier. Pour cela, nous considérons la matrice carrée[1] appelée *univers des probabilités de cooccurrences*, dont les lignes et colonnes correspondent aux situations. Ses lignes fournissent pour chaque situation $s=e_t$ son taux de cooccurrences avec chacune des situations $s'= e'_{t'}$ (pondéré par l'éloignement dans le temps) et sont normalisées en profils. Les composantes des profils lignes $P_s^{s'}$ ainsi constitués sont :

- Nulles[3] si $t>t'$
- si $t<=t'$, le produit de trois facteurs :

$$P_s^{s'} = \alpha_s \beta_s^{s'} \Phi_s^{s'}$$

[1] Dans la présentation où on ne considère que les transitions vers le futur, la matrice est en fait une matrice carrée triangulaire droite. Néanmoins, on peut choisir d'intégrer les transitions venant du passé.

- Le terme Φ mesure la fréquence relative des individus dans la situation *s* qui connaîtront la situation *s'* (probabilité empirique de connaître à la fois les situations *s* et *s'*):

$$\Phi_s^{s'} = \frac{\sum_{i=1}^{I} Y_i^s Y_i^{s'}}{\sum_{i=1}^{I} Y_i^s}$$

Le coefficient de pondération β, décroît avec le délai (*t'*-*t*). Cette fonction est au libre choix de l'utilisateur. Dans notre exemple, nous avons choisi arbitrairement l'inverse du délai :

$$\beta_s^{s'} = \frac{1}{t'-t+1}$$

- Le coefficient α, assure que la somme en ligne des probabilités est constante et égale à 1. Sans le paramètre α, la somme en ligne varierait en raison de la nullité des probabilités de transition vers le passé et de la censure à droite:

$$\alpha = (\sum_{s'} \beta_s^{s'} \Phi_s^{s'})^{-1}$$

Les lignes du tableau sont donc normalisées comme des profils (la somme des lignes vaut 1).

La distance entre deux situations *s* et *s'*est donc définie naturellement comme la distance du χ^2 entre profils-lignes P_s et $P_{s'}$. Entre autres propriétés, elle répond au principe d'équivalence distributionnelle.

$$d^2(s,s') = \sum_{\sigma=1}^{p} \frac{1}{P_\cdot^\sigma} \left(\frac{P_s^\sigma}{P_s^\cdot} - \frac{P_{s'}^\sigma}{P_{s'}^\cdot} \right)^2 = \sum_{\sigma=1}^{p} \frac{1}{P_\cdot^\sigma} \left(P_s^\sigma - P_{s'}^\sigma \right)^2$$

Remarques sur la cohérence de l'approche:

- Le taux moyen de transitions est une grandeur et suit asymptotiquement une loi normale. Il s'inscrit naturellement dans une approche euclidienne.

- A un instant donné, un état est caractérisé par un profil de transitions vers le futur (le barycentre des transitions des individus qu'il regroupe). Cette approche est en cohérence avec le pré-codage de l'information qui a aussi regroupé les états en tenant compte de ce critère : le CDI regroupe le « vrai » CDI et les fonctionnaires sur un critère de stabilité. Les transitions du groupe *CDI et fonctionnaires réunis* est le barycentre des transitions des deux groupes *CDI* et *fonctionnaires* séparés dans des proportions identiques à celles de leur fréquences respectives.

- Le profil de transitions défini comme barycentre des individus qu'il regroupe est en cohérence avec notre choix futur de considérer qu'une classe est représentée par le barycentre des individus qu'elle rassemble, et de l'illustrer par son chronogramme. Le chronogramme a en effet la double propriété de caractériser les individus et de donner les poids des barycentres des états à chaque instants.

Les transitions des chômeurs sont les transitions du groupe des chômeurs et de la même façon, l'état moyen entre chômeur et inactif est l'ensemble des transitions d'un groupe constitué d'autant de chômeurs que d'inactifs. En effet, les transitions d'un groupe d'individus entre deux dates sont les sommes des transitions individuelles. Et de la même façon, les transitions d'une population (les chômeurs) identifiée comme un mélange de deux catégories d'individus (chômeurs indemnisés et non indemnisés, femmes et hommes) sont les combinaisons linéaires des transitions des catégories, les coefficients de la combinaison étant ceux du mélange. Cette approche est en cohérence avec un algorithme de regroupement.

Elle est aussi en cohérence avec le pré-codage des états dans le calendrier. Ces états sont des regroupements d'états élémentaires (l'état CDI regroupe le CDI et les fonctionnaires) faits à partir d'un critère de stabilité ou

de précarité en regard à la qualité des transitions. Si on regroupe des *états élémentaires* (*emploi jeune*, *CES*, *CIVIS*) en un seul *état général* (*contrat aidé*), ce dernier peut être considéré comme un mélange des états *élémentaires* dans les proportions correspondant aux contingences et variant dans le temps avec elles. Les transitions entre deux dates depuis (resp. vers) cet état *général* sont alors exactement les combinaisons linéaires des transitions depuis (resp. vers) les états *élémentaires*, et les coefficients de cette combinaison linéaire sont ceux des proportions du mélange. Il y a donc une correspondance linéaire entre deux niveaux de description emboîtés des états quand on travaille avec les transitions.

L'approche euclidienne est donc d'avantage adaptée aux propriétés linéaires des transitions qu'une approche discrète par exemple.

Dans une approche discrète (comme dans l'optimal matching), chaque état est différent de l'autre et la moyenne entre deux états n'a pas de sens : on est soit *chômeur*, soit *inactif* mais pas entre les deux. L'état moyen entre *chômeur* et *inactif* n'a pas de sens puisqu'il n'y pas de définition de cet état (l'état *hors emploi* n'étant pas dans la liste des états).

Remarques techniques :

- Pour toute situation *s* et pour toute échéance future *t'=t'*$_0$, *la somme sur les s' des transitions pour l'échéance t' vaut 1* ce qui équivaut à « *la somme des E valeurs de* $\Phi_s^{s'}$ *vaut 1* ». Chaque individu passant par *s* passe nécessairement par un et un seul état en *t'* : $\forall\, s,\quad \sum_{t'=t'_0} \Phi_s^{s'} = \mathbf{1}$

- Pour les mêmes raisons, dans le cas où β ne dépend que du délai *t'-t*, α ne dépend que du nombre d'instants futurs renseignés : $\alpha = (\sum_{s'} \beta_s^{s'})^{-1}$

2.2. La structure de l'espace des situations, sa dynamique temporelle et la définition des évènements principaux

Une fois déterminée la distance inter-situations, on peut en déduire la structure de l'espace des situations. Le procédé très classique, qui repose sur la formule de Torgerson (Torgerson 1958, Benzécri 1973), nous fournit la matrice des produits scalaires entre situations à partir de celle des distances inter-situations. La diagonalisation de cette matrice fournit les composantes principales du nuage des situations. Nous appellerons ***évènements principaux*** les composantes principales du nuage des situations.

Par la suite, La décomposition linéaire des situations sur les évènements principaux, décrite plus haut, permet l'écriture de la trajectoire dans l'espace de ces évènements principaux. La distance entre les trajectoires est la distance euclidienne dans cet espace. Elle peut donc être utilisée par toute méthode de classification de type euclidienne comme les centres mobiles, les classifications avec le critère de Ward ou les cartes d'auto-organisation avec l'algorithme de Kohonen qui sera présenté plus loin.

$\Delta_s^{s'} = -\frac{\mathbf{1}}{\mathbf{2}}\left[d^{ss'} - d^{s\cdot} - d^{\cdot s'} + d^{\cdot\cdot}\right]$, où $\Delta_s^{s'}$ est la matrice des produits scalaires entre les situations *s* et *s'* et $d^{ss'}$ la distance entre *s* et *s'*

Pour montrer la capacité de cette méthode à prendre en compte la structure entre les états de travail et leur dynamique dans le temps, on montre ici que la corrélation entre les situations intègre que le *CDD* a un statut qui favorise l'insertion vers le *CDI* en début de période puis le perd pour devenir par la suite un statut précaire au même titre que *l'intérim*.

Pour l'illustrer, on présente dans les figures 4 et 5 le suivi dans le temps de la corrélation entre les situations à instants identiques. La figure 4 montre la série dans le temps des corrélations au *CDI*. L'abscisse est la

date, l'ordonnée la corrélation entre la situation de *CDI* à cette date et chacun des trois situations *CDD*, *chômage* et *intérim* à cette même date. Les corrélations entre une situation de *CDI* à la data *t* et une situation de *CDD* ou *chômage* à une date $t+\tau$ ne sont pas représentées dans cette figure mais en figure 6. La corrélation entre *CDI* d'une part et *CDD*, *chômage* et *intérim* d'autre part décroit dans le temps pour devenir très faible. Ceci correspond au fait que les trajectoires se stabilisent en fin de période. Cette figure montre également que la décroissance se fait par palier. Pour le *CDD*, on constate trois paliers d'insertion vers le *CDI* : forte (jusque mi-2001), moyenne (jusque début 2003) et faible (au-delà). Pour l'intérim et le chômage, on n'en trouve que deux : moyenne (jusque mi-2001) puis faible. La figure 5 complète cette structure temporelle en montrant les corrélations au *chômage* pour des situations de même date. On constate un maintien de corrélation au *chômage* de niveau moyen du *CDD* et de *l'intérim* en phase finale, ce qui confirme le statut de précarité de ces états pour cette phase. En début de période (jusque mi-2001), les corrélations entre les situations de *chômage* et *CDD* sont fortes du fait qu'elles cumulent simultanément des propriétés d'insertion vers de *CDI* et de précarité. Ensuite, dans une deuxième phase, seule la précarité les rapproche. Dans cette deuxième phase, la figure 5 montre que *l'intérim* et *CDD* jouent un rôle similaire vis-à-vis du chômage.

En conclusion, le *CDD* joue deux rôles –d'insertion au *CDI* et de précarité- selon trois phases la première où *l'insertion est dominante*, la seconde où les *deux rôles sont présents* et la troisième où *la fonction d'état précaire prend le pas sur celui d'insertion*. De plus, dans la phase de précarité, les corrélations du *CDD* et de *l'intérim* au *chômage* sont identiques, montrant leurs rôles similaires.

La figure 6, quant à elle montre la corrélation entre les situations de *CDI* avec les situations antérieures de 12 mois. Cette figure confirme celles des deux autres figures en confirmant que la corrélation du *CDI* avec lui-même (12 mois plus tôt) est la plus forte

Remarque :

- La corrélation pour notre métrique indique si les situations ont des profils de transitions probables vers des situations futures voisines. La dynamique est plus judicieuse à interpréter que le niveau dans la mesure où ce dernier dépend du paramètre, ici arbitraire, β.
- L'évolution décroissante de la corrélation entre *CDD* et *CDI* en trois paliers sera reprise et confirmée lors de l'application 3.2.3.

Figure 4

Evolution de la corrélation au *CDI* pour des situations de mois identiques et des états de *CDD* (2), *intérim* (5), *chômage* (6)

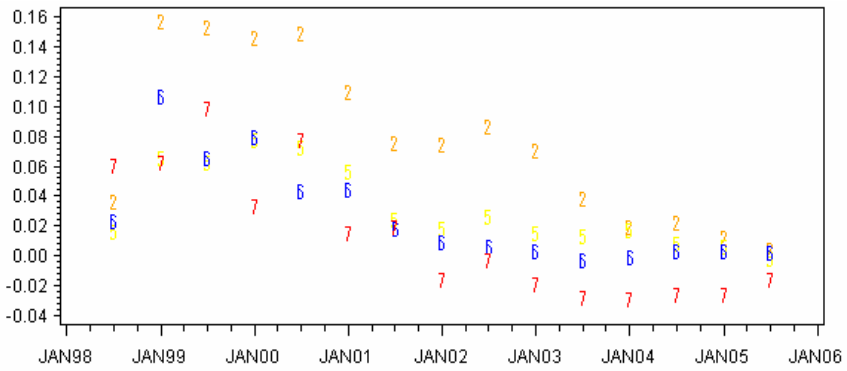


Figure 5

Evolution de la corrélation au chômage pour des situations de mois identiques et des états de *CDI (1)*, *CDD* (2), *intérim* (5)

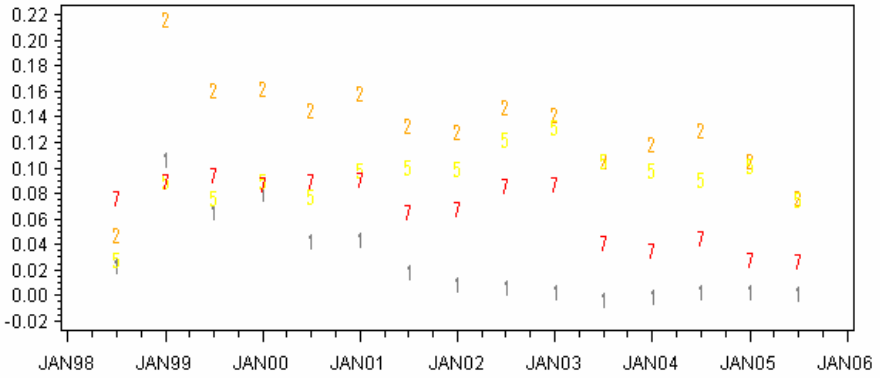
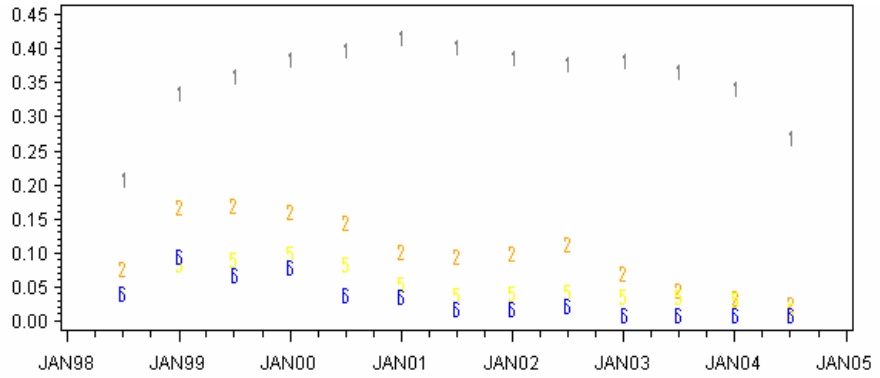


Figure 6

Evolution de la corrélation au CDI pour des situations antérieures d'un an et des états de *CDI (1)*, *CDD* (2), *intérim* (5), *chômage* (6)



2.3. Les données manquantes

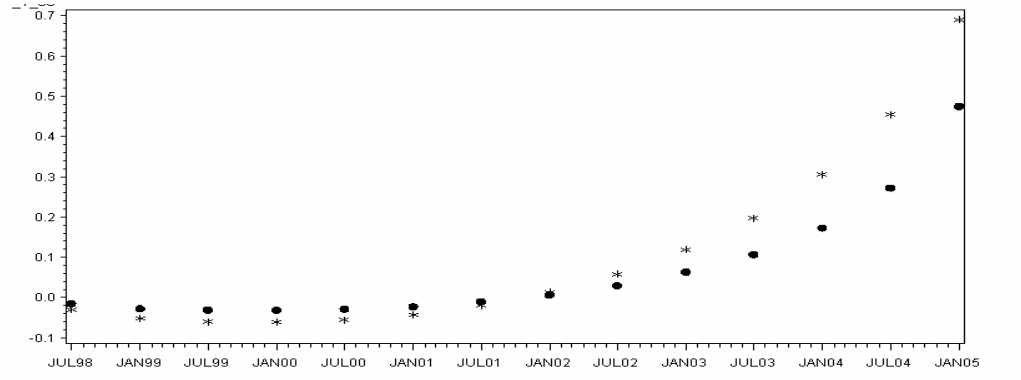
La question des données manquantes est une question récurrente dans le cadre de l'analyse de trajectoires. En particulier, le problème des censures à droite ou à gauche est souvent structurel au recueil des données, par exemple lorsque certains individus ont vécu ou travaillé plus longtemps que d'autres. C'est pourquoi nous y consacrons une section alors que dans l'application sur l'insertion, les données ne souffrent pas de valeurs manquantes.

Il y a deux stratégies complémentaires pour aborder cette question. On peut déduire les données manquantes de la logique de la trajectoire de l'individu elle-même (l'individu s'inscrit dans un parcours stable de *CDI* ou d'exclusion du marché du travail par l'inactivité). On peut aussi les déduire de la logique de parcours d'un ensemble de trajectoires (toutes les trajectoires à priori, les trajectoires de la classe à postériori). Notre méthode cumule ces deux approches : elle permet de projeter les trajectoires sur les évènements principaux qui recouvrent toute la période et donc de les projeter sur les parties non renseignées. Dans le cas particulier de la censure à droite, la matrice des corrélations fournit précisément une projection des situations sur leur futur et donc la recodification des trajectoires réduites se fait également sur la période totale par projection sur le futur. La période correspondant aux données manquantes aura néanmoins moins de poids à mesure de

son éloignement dans le temps des occurrences renseignées. Cette décroissance est contrôlée par le paramètre β. La figure 7 illustre les corrélations entre la situation s'=(*CDI,T*) de *CDI* à l'instant final et les situations s=(*CDI,t*) de *CDI* aux dates antérieures pour les fonctions $\beta(s,s')$ respectives $1/(1+0.3*(t-t'))$ et $1/(1+(t-t'))$. Pour la première le coefficient de corrélation (0.68 à 6 mois, 0.45 à 12 mois et 0.3 à 18 mois) est similaire à la seconde avec un décalage de six mois (0.48 à 6 mois, 0.27 à 12 mois et 0.18 à 18 mois). La première fonction projette donc avec le même poids six mois plus loin que la seconde. Le paramètre β permet ainsi de contrôler l'horizon de projection pour un niveau de corrélation donné.

Figure 7

Corrélation entre le CDI en position finale et en CDI à une date précédente pour la fonction * : $\beta=1/(1+0.3*t)$ et ● : $\beta=1/(1+t)$



3. LA MÉTHODE D'AGRÉGATION DES CLASSES

Ayant défini la distance entre trajectoires à classer, nous nous intéressons à présent dans cette section à la procédure d'agrégation en classes. D'après ce que nous venons de voir, les algorithmes adaptés aux distances euclidiennes sont tous des candidats potentiels, aucun n'étant par ailleurs spécifique à la distance qui nous intéresse. Nous présenterons d'abord le couplage *centres mobiles - classification hiérarchique*, qui est la méthode la plus utilisée en classification des trajectoires, puis dans un second temps, les cartes d'auto-organisation.

3.1. Le couple centres mobiles – classification hiérarchique

Partant des 16 000 trajectoires individuelles, nous définissons d'abord 50 classes avec une méthode d'agrégation autour de centres mobiles, qui sont ensuite agrégées en 8 classes au moyen d'une classification hiérarchique (avec le critère de Ward) en intégrant le poids des classes.

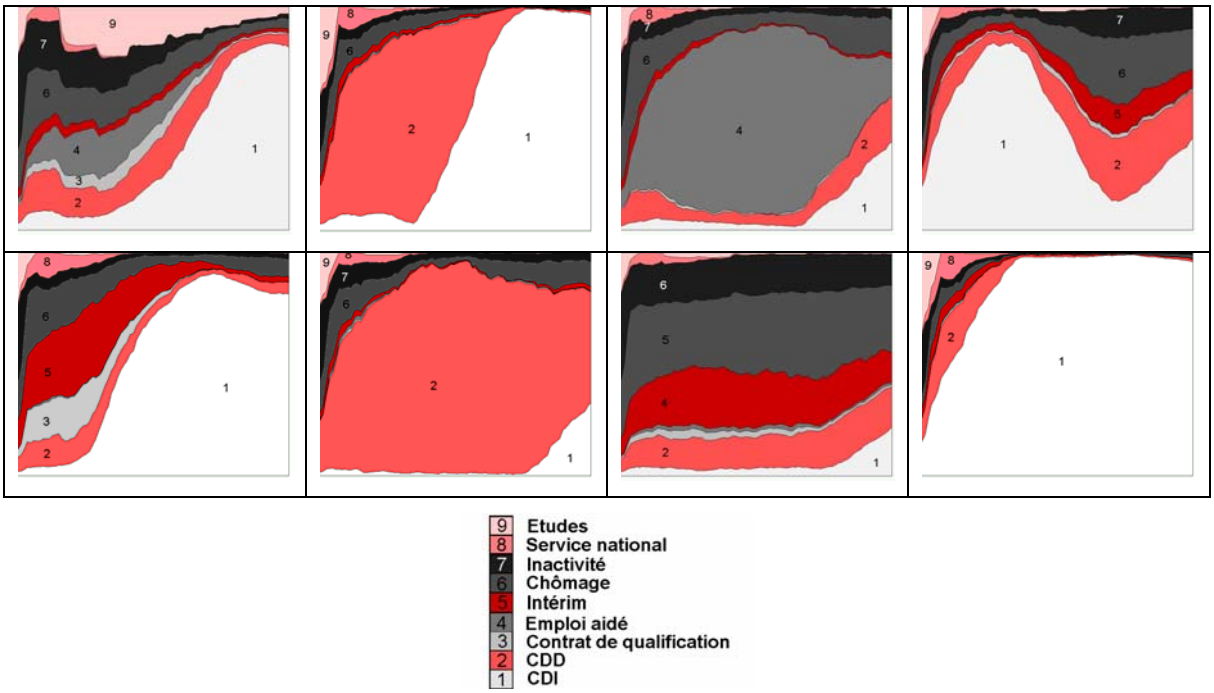
Le niveau de partitionnement en 8 classes montre l'intérêt de la distance qui met en évidence la dynamique temporelle. La classification distingue en effet des types de trajectoires stables (*de précarité* ou *dans l'emploi*) et des trajectoires qui incluent des dynamiques temporelles *d'insertion vers l'emploi* ou *de perte de CDI*. Il montre aussi la prise en compte des proximités en fin de parcours entre *l'intérim*, *le CDD* et *le chômage*. La contribution commune de ces trois états a permis de faire émerger dès le niveau de partition en 8 classes des trajectoires « descendantes » (*perte de CDI*) et de regrouper (en classe 7) des trajectoires de précarité hors emploi avec des trajectoires *d'allers et retours entre le chômage et les contrats courts*. Le type de trajectoires « descendantes » est important d'un point de vue économique et social car il montre la vulnérabilité de jeunes occupant un emploi réputé stable et est, de plus, atypique par sa dynamique temporelle. Ce résultat est donc largement conforme à nos objectifs.

Une classification moins agrégée en 16 classes montre l'intérêt d'affiner la partition et plaide pour un dispositif de regroupement facilitant l'analyse d'un grand nombre de classes tel que les cartes d'auto-organisation (cf. infra). La partition en 16 classes détaille le niveau supérieur de façon très satisfaisante. En termes de significativité, les effectifs sont conséquents. En termes de signification, l'accès à moyen terme au *CDI* (au niveau 8 : *CDD->CDI* et *autres->CDI*) est détaillé par catégorie (*contrats de qualification*, *intérim*, *chômage*). Les états de *chômage* ou d'*inactivité* quasi-permanents émergent de la classe *hors-emploi*. Par ailleurs, la dynamique temporelle des *contrats aidés* apparaît.

Figure 8

Typologie de l'insertion des jeunes sortis de la formation initiale en 1998 sur 7 ans

Partition en 8 classes à partir du couple de classification centres mobiles – classification hiérarchique



3.2. Les cartes d'auto-organisation

Les cartes d'auto-organisation, qui utilisent l'algorithme non supervisé de Kohonen (Kohonen, 2001 ; Cottrell et al., 2003; Fort 2006), ont souvent été utilisées en analyse des données (Oja et Kaski 1999, Rousset et Guinot 2002) y compris pour des données longitudinales (Cottrell et al 1998, Giret et Rousset 2005, Delaunay et Lelièvre 2006, Massoni et al 2009). La méthode de classification à l'aide des cartes d'auto-organisation appartient à la famille des algorithmes conduisant à des partitions, comme les méthodes d'agrégation autour des centres mobiles (Lebart et al 2006). Cette méthode en partage des points communs : le nombre de classes est fixé par l'utilisateur, et elle est adaptée aux données de grande taille. Son originalité est d'introduire une notion de voisinage entre les classes et d'utiliser un support graphique associé, appelé « carte », qui organise les classes par proximité. Cette caractéristique permet de travailler sur un grand nombre de classes, travail souvent fastidieux avec les méthodes classiques (partitions emboîtées rapidement peu lisibles). Au delà de la proximité des classes prises deux à deux, l'ordonnancement plus global d'un ensemble de classes peut être interprété à partir d'une dimension endogène ou exogène. C'est le cas, par exemple, si l'ordre des classes correspond à l'ordre croissant de cette dimension (le temps) ou à l'importance du retard au déclanchement d'un évènement (comme l'obtention d'un *CDI*). Sa capacité à traduire visuellement l'évolution temporelle rejoint ainsi les objectifs de l'analyse longitudinale.

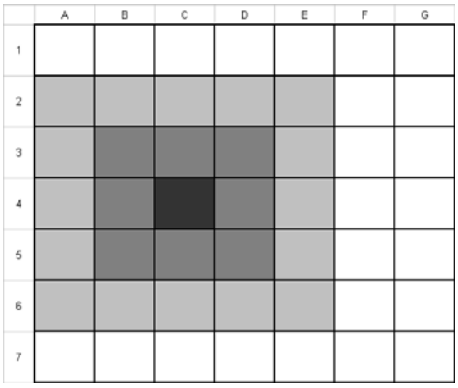
3.2.1. Présentation des CAO

La carte est un réseau dont le nombre d'unités et la structure sont choisis par l'utilisateur. Chaque unité est représentée par un rectangle qui symbolise une classe. Les structures les plus usuelles sont la ficelle unidimensionnelle et la grille bidimensionnelle (figure 9). La ficelle est généralement utilisée lorsque l'on désire faire apparaître un ordre entre les classes, alors que la grille met en valeur des liaisons inter-classes multidimensionnelles.

La notion de voisinage est induite par la structure de la carte. Sont considérées comme voisines, les unités proches sur la carte (*cf.* ci-après). Le résultat de l'algorithme est d'affecter à chaque unité un vecteur de l'espace des données appelé « vecteur code ». A chaque individu de l'espace des données, on attribue l'unité dont le vecteur code est le plus proche. On constitue ainsi une classification. Les vecteurs code convergent vers les centres de classe respectifs (les barycentres des classes). La propriété majeure de l'algorithme est que deux individus associés à des classes, dont les unités sont voisines sur la carte, sont voisins dans l'espace des données. Sur la figure 9, les unités voisines de celle numérotée C4, aux rayons 2, 1 et 0, sont respectivement les unités grisées (rectangle A2-E6), les seules unités grisées foncées (rectangle B3-D5), et elle-même.

Figure 9

Carte d'auto-organisation : exemple de grille bidimensionnelle



Légende : Les unités voisines de la classe C4 gris foncé au rayon 1 sont les unités gris moyen (rectangle B3-D5) ; au rayon 2 s'ajoutent celles grisées clair (rectangle A2-E6). Les individus associés à des classes voisines sont proches dans l'espace des données. Les individus des classes C4 et C5 présentent donc des trajectoires professionnelles avec de grandes similitudes.

Les cartes d'auto-organisation comme outils graphiques d'analyse

Les unités de la carte, matérialisées par des rectangles, sont considérées comme des fenêtres graphiques. Elles permettent de représenter l'information désirée sur toutes les classes simultanément (Rousset et Guinot 2002, Cottrell et al 2003). On peut ainsi lister le nom des individus de chaque classe comme en analyse en composantes principales, y inscrire une représentation des trajectoires à l'aide de chronogrammes (*cf.* section 1 pour une définition) ou représenter une dimension qualitative par un camembert ou un histogramme. Le grand avantage de cette représentation provient de l'organisation des unités, c'est-à-dire des fenêtres, par voisinage. Elle permet de traiter ensemble les classes voisines qui ont une caractéristique commune. On parle alors de la caractéristique d'une région de la carte (zone connexe de la carte). De la même façon, elle indique qu'une caractéristique est partagée par deux populations, par ailleurs globalement différentes, lorsqu'elle se retrouve sur deux régions éloignées sur la carte. L'analyse par régions de voisinage permet de traiter une grande partie de la redondance. En effet, dès l'instant où deux classes sont voisines, on s'attend à ce qu'elles aient un grand nombre de caractéristiques communes du point de vue des variables endogènes et exogènes. Dans le cas de trajectoires, la même évolution décalée dans le temps génère une redondance qui peut être traitée de la sorte. L'ensemble des caractéristiques communes s'interprète alors par le fait que l'on a deux populations d'individus très proches. Cette méthode met en évidence des effets locaux (entre classes voisines); c'est-à-dire, les dimensions qui éloignent ou rapprochent des classes voisines.

3.2.2. Application des CAO: une typologie des parcours d'insertion

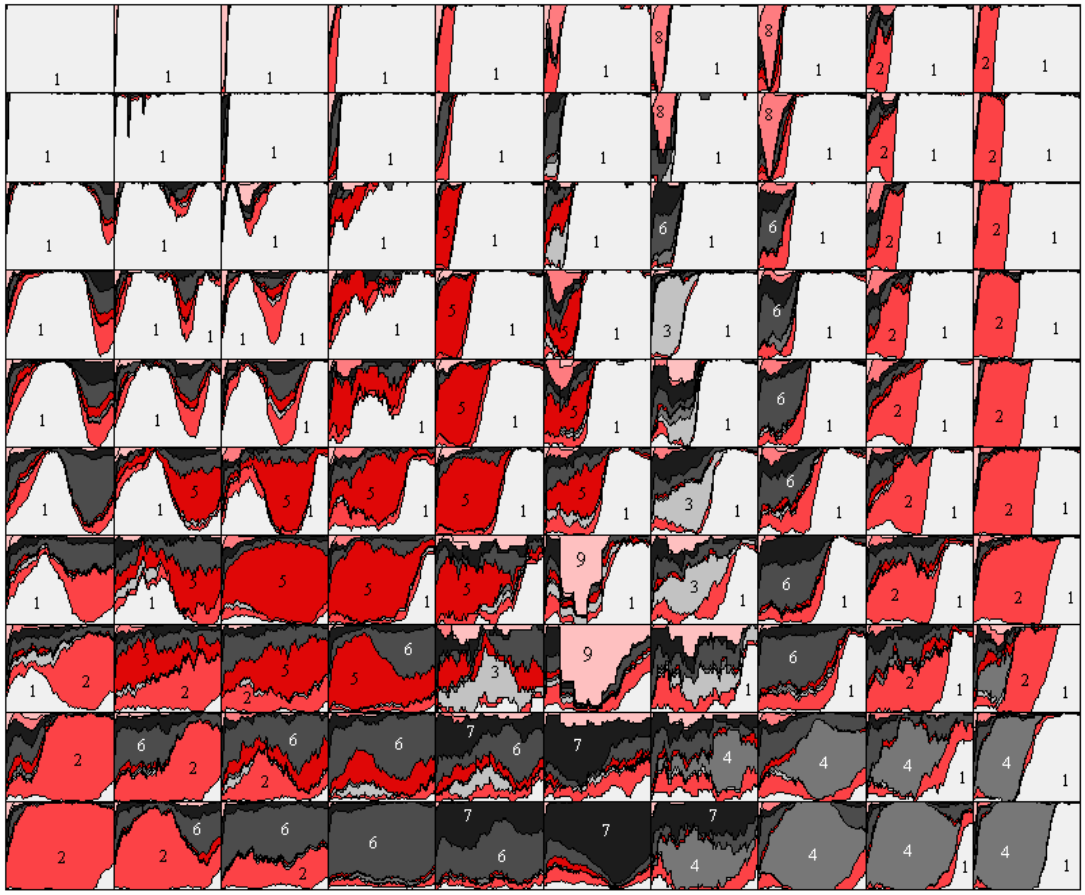
On établit une typologie des parcours à partir des cartes d'auto-organisation appliquée avec la distance construite en section 2. Les réseaux choisis sont une grille 6x6 (*cf.* annexe 1) et une grille 10x10. La grille 6x6, plus petite, sert à illustrer la méthode (annexe 1). La grille 10x10 est le niveau qui nous a servi à l'exploitation dans notre application. La carte 1 représente la solution de l'algorithme, chaque case est caractérisée par le chronogramme de sa cohorte.

La carte de Kohonen permet d'avoir une vision relativement synthétique des différents types d'itinéraires qui caractérisent les 7 premières années de vie active. Elle fait également apparaître le voisinage entre les différents parcours. Globalement, la carte est dominée par une opposition Nord/Sud entre des trajectoires de stabilisation dans des emplois à durée non limitée et des trajectoires de précarité dans l'emploi ou hors emploi. Dans la partie Nord de la carte, l'accès à un emploi stable (contrat à durée indéterminée, fonctionnaires et indépendants) se fait rapidement. Dans la partie Nord-Ouest, principalement dans les trois premières colonnes des deux premières lignes, l'accès à ce type d'emploi est immédiat et pérenne. Progressivement, lorsque l'on se déplace vers la partie Est de la carte, une période transitoire apparaît en début de trajectoire, dominée par le service militaire ou des emplois à durée déterminée. En descendant vers le Sud de la carte apparaissent graduellement les trajectoires sans passage par l'emploi à statut stable. Dans la partie Est de la carte, cet aspect se cristallise sur les types de contrats à durée limitée, les jeunes se maintenant en emploi tout au long de leur sept premières années de vie active. La partie Centre-Ouest décrit des trajectoires de cessation de *contrats à durée indéterminée*. Elle les décline selon le type de contrat qui a suivi. La partie Sud est découpée en trois : à l'Ouest les parcours de maintien en emploi à durée déterminée, au centre le hors-emploi (*chômage* et *inactivité*) et à l'Est les contrats aidés.

Carte 1

Typologie de l'insertion des jeunes sortis de la formation initiale en 1998 sur 7 ans

Classification obtenue à l'aide d'une cartes d'auto-organisation 10x10



Chaque cellule correspond au chronogramme d'un groupe de trajectoires où est décrite la situation mensuelle de l'ensemble des individus de ce groupe. Le code couleur est le suivant :

3.2.3.L'apport des CAO pour l'analyse longitudinale

Nous allons dans cette section illustrer l'apport des cartes d'auto-organisation en montrant d'une part nombre de propriétés que l'on pouvait anticiper et d'autre part une application de ces propriétés.

- Le système de représentation des CAO, organisées par proximité des classes, facilite la lecture simultanée d'un grand nombre de classes et de leur connexité. Deux arguments plaident pour un grand nombre de classes. D'abord, dans le cadre longitudinal, il y a souvent une grande diversité de trajectoires. Dans notre exemple, il existe potentiellement 9^{88} trajectoires et parmi les 16000 trajectoires observées, 12000 sont différentes (parmi les autres, 3500 sont des « *CDI* tout au long du parcours »). Ensuite, dans le cas de situations fréquentes (par exemple en raison d'états

- La proximité des classes sur la carte se traduit par une similitude des chronogrammes, grâce à la propriété de conservation de la topologie des cartes d'auto-organisation (deux individus de classes voisines sont proches dans l'espace des données). On peut noter d'une part que les nombreuses classes correspondant à des variantes d'accès rapide au *CDI* sont regroupées au Nord de la carte. D'autre part, qu'un décalage dans le temps d'une même évolution va se lire dans une série de classes voisines, tel le passage de *CDD* à *CDI* qui se fait de plus en plus tardivement au fur et à mesure qu'on descend le bord droit de la carte. Enfin, deux situations proches se retrouvent proches sur la carte, tout comme leurs évolutions (*chômage* et *inactivité*, *intérim* et *chômage*, selon une intensité qui varie avec le temps) ce qui permet de traiter simultanément les proximités entre états et temporelles. Ainsi les propriétés des classes peuvent être généralisées à leur région d'appartenance comme une variation ou une évolution d'un même événement. La carte fournit un niveau de regroupement à 100 classes mais permet de l'interpréter par régions. Le découpage de la carte en régions est au libre choix de l'expert. Rousset et Guinot (2002) proposent d'utiliser une classification supplémentaire sur les représentants de classes pour le réaliser).

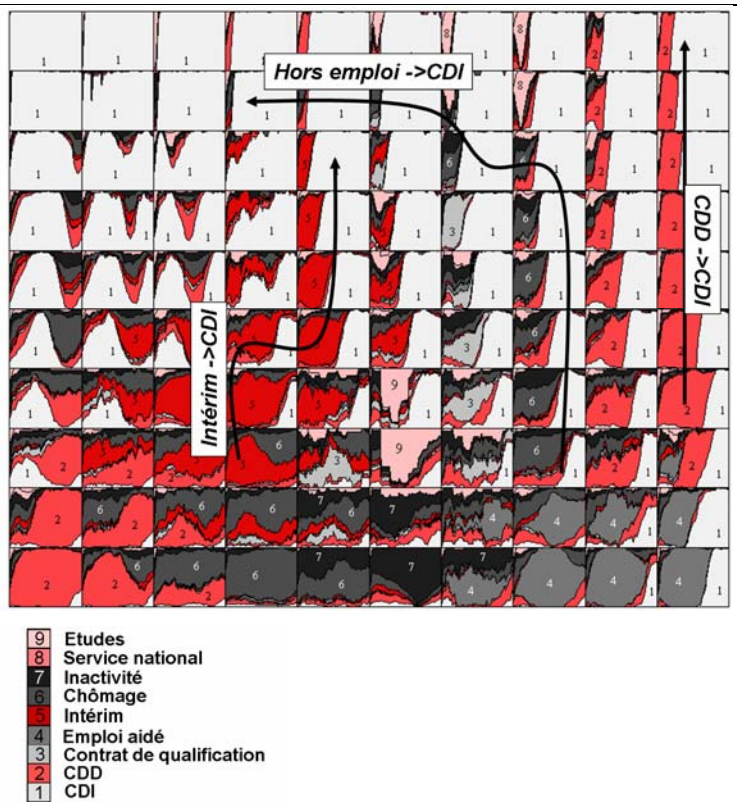
Illustrons l'apport des CAO par sa capacité à mettre en évidence le *pouvoir d'intégration* des contrats à durée limitée. La carte (carte 1) oppose deux types de processus : celui de « contrat intégrateur » vers les contrats à durée indéterminée (Centre-Est) et celui de « trappe à emploi précaire » (Sud-Ouest).

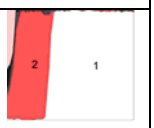
Les deux colonnes de droite (carte 1) *I* (I8 à I1) et *J* (J8 à J1) décrivent les vagues d'accès au *CDI* via le *CDD* et les ordonnent dans le temps. Par rapport à la colonne *J*, La colonne *I* est constituée de classes d'individus qui ont eu moins recours au *CDD* en intensité pendant la période pré-*CDI*. Le film de ces flux sont représentés tableau 2 colonne de droite nommée *zoom*. Il est possible à partir des effectifs des classes et en mesurant le décalage temporaire que constitue une classe par rapport à la précédente de déduire une évolution des flux respectifs d'accès au *CDI*. Le décalage dans le temps est mesuré de la façon suivante : pour chaque classe, on repère le mois m_{85} où 85% de la cohorte de la classe a atteint de *CDI* (les pentes des chronogrammes étant, à ce moment, très grandes, la mesure est assez fiable). Le flux moyen est défini comme le nombre d'accès (effectif de la classe) rapporté à l'écart dans le temps. L'évolution des flux ainsi constituée montre une première phase, avant novembre 2000, où le potentiel d'insertion du *CDD* vers le CDI est fort (niveaux de 57,6 et 57). Le potentiel intégrateur du *CDD* tend ensuite à se stabiliser jusqu'en juillet 2003 à 50 % de la phase précédente (29, 30,4 et 27,5). Cette évolution confirme l'évolution décroissante de la corrélation entre *CDD* et *CDI* en trois paliers (avant juillet 2000, entre juillet 2000 et 2003 et au-delà) décrite en section 2.2.

Cette application illustre l'intérêt des cartes d'auto-organisation dans la phase d'analyse des chronogrammes. Leurs propriétés d'organisation dans le temps et de constitution de populations significatives et homogènes permettent de travailler à un niveau fin de description avec un support graphique.

Figure 10

Vagues d'intégration au *CDI* à partir du *CDD* : détail de la colonne 10



Effectif m_{85}	Ecart temporel Flux	zoom
E= 398 m_{85}=01/2000	Ec. tp. =. F.=.	
E= 288 m_{85}=06/2000	Ec. tp. =5 F.=57.6	
E= 285 m_{85}=11/2000	Ec. tp. =5 F.=57	
E= 163 m_{85}=07/2001	Ec. tp. =8 F.=20.4	
E=173 m_{85}=01/2002	Ec. tp. =6 F.=29	
E= 243 m_{85}=09/2002	Ec. tp. =8 F.=30.4	
E= 248 m_{85}=07/2003	Ec. tp. =9 F.=27.5	

4. MISE EN PERSPECTIVES AVEC QUELQUES MÉTHODES DE CLASSIFICATION USUELLES DANS L'ANALYSE LONGITUDINALE

Il faut d'abord s'entendre sur ce que l'on appelle méthode de classification : il s'agit pour nous de l'ensemble du processus : (re)codage des données (qui sélectionne ou condense l'information définissant les éléments à classer), choix de la mesure de distance entre éléments à classer, choix de l'algorithme d'agrégation. Comme toujours en traitement de données, le choix de la méthode est conditionné par les propriétés de ces données que l'on a choisi d'analyser. Dans notre cas, c'est *la discrimination entre les trajectoires induite par l'écart entre les états et leur évolution dans le temps*. Il convient donc d'examiner comment les méthodes intègrent ce critère. Nous privilégions ici trois méthodes souvent utilisées dans les analyses longitudinales.

4.1. Distance du χ^2 et classification des calendriers

La classification se fait directement sur le tableau des calendriers – ou de l'ensemble des situations- avec la distance du $\chi 2$ (la plupart des programmes demandent que le tableau soit mis sous forme disjonctive-complète, cf. section 2.1). Le plus souvent, on mixte Classification ascendante hiérarchique et agrégation autour de centres mobiles, avec le critère d'agrégation de Ward. Fréquemment aussi, on réalise au préalable une ACM des données, afin d'éliminer les bruits portés par les facteurs de rang élevé et stabiliser le résultat de la classification, qui est faite alors sur les premiers facteurs (pris en nombre suffisant pour totaliser une part importante de l'inertie). Avec cette méthode, l'ordre des séquences n'est pas pris en compte directement. Il l'est cependant de façon indirecte par le biais de la distance du $\chi 2$ qui pondère les écarts entre états par l'inverse de leur fréquence, et fait donc émerger la dynamique temporelle, par exemple parce que certains états sont rares au début de la phase d'insertion et communément partagés à la fin. Les modalités de faible effectif, pesant assez lourd dans la distance du $\chi 2$, peuvent apparaître à un niveau très haut d'agrégation. Dans le cas où les situations peu fréquentes sont d'un moindre intérêt du point de vue de l'étude, comme par exemple le *service national* après la fin de la conscription, il est alors nécessaire de sélectionner, dans l'arborescence, les branches les plus significatives ou les plus intéressantes. La production d'un grand nombre de classes à faibles effectifs peut ainsi apparaître comme un inconvénient de cette méthode, lorsqu'elles deviennent trop encombrantes. Pour un regroupement à l'aide des cartes, on pourra se référer à Delaunay et Lelièvre (2006).

4.2. L'optimal matching

Cette méthode se distingue par le choix de la distance entre trajectoires. La technique, issue de l'analyse génétique, consiste à comparer deux chaînes $a=(a_1,\ldots, a_m)$ et $b=(b_1,\ldots,b_n)$ en mesurant un coût de transformation de l'une à l'autre à partir de trois opérations élémentaires : la substitution, l'insertion et la suppression. Chacune étant affectée d'un coût élémentaire, le coût de transformation de a en b est la somme des coûts élémentaires.

Ce calcul de la distance entre trajectoires est très utilisé dans les pays anglo-saxons pour analyser les carrières professionnelles (Abbott et Hrycak, 1990 ; Halpin et Chan, 1998) ou les parcours d'insertion (Scherer, 2004 ; Brinsky Fay, 2007). Contrairement aux distances précédentes, elle considère l'espace comme discret. Elle ne se soucie pas de la cohérence entre les différentes transformations de la même trajectoire a selon qu'elle est comparée à la trajectoire b ou c. La question de la mesure des coûts élémentaires et de leur pertinence par rapport à la question posée fait l'objet de débats au sein de la communauté des sociologues (Wu, 2000). Par rapport à notre problématique, les coûts élémentaires sont constants dans le temps, ils ne peuvent donc prendre en compte une évolution des rôles d'insertion des différents états. Par rapport à l'exploitation des transitions, plusieurs arguments s'opposent à les utiliser pour

mesurer les coûts élémentaires, citons en deux. Le premier reprend la remarque de la section 2.1. Les transitions sont des grandeurs quantitatives qui suivent asymptotiquement des lois normales et donc évoluent dans un espace euclidien. Il est donc contre productif de travailler avec des matrices de dissimilarité ou avec une distance de type Hamming. Le second est de type technique : si l'on estime qu'il est judicieux de transformer les trajectoires transformées à l'aide des opérations d'insertion et de suppression, il convient de mesurer les transitions à partir des trajectoires transformées. Or les coûts étant calculés avant, cet ordre ne peut être respecté. Travailler seulement avec les couts de substitution résoudrait ce problème technique mais, l'optimal matching se réduirait alors à la distance de Hamming qui ne tient pas compte de l'ordre séquentiel, ce qui serait en contradiction avec une approche par les transitions.

4.3. L'analyse harmonique

L'analyse harmonique (Deville, 1977, Robette et Thibault, 2008) se distingue par le recodage, préalable à la classification, qui condense l'information portée par les données, pour en réduire la complexité ou parce que l'unité de temps pertinente n'est pas celle du calendrier de départ. On découpe la période couverte par le calendrier en fenêtres de temps (par exemple dans le cas du calendrier mensuel, on regroupe en années). Ensuite, pour chaque individu et chaque fenêtre, on mesure la durée passée dans chaque état. La distance est à nouveau la distance du $\chi 2$. Cette méthode essaye de traiter ensemble la mesure inter-états et inter-individus en utilisant un recodage de l'information. Ce qui rapproche deux états, c'est leur présence dans une même fenêtre de temps pour un même individu. Ce recodage demande un certain nombre de concessions. Tout d'abord une perte d'information, l'unité de temps est élargie très significativement (les fenêtres doivent agréger un grand nombre d'unités de temps pour que l'opération ait un sens). L'ordre des séquences à l'intérieur d'une fenêtre est perdu : une transition *CDD->CDI* est comptabilisée de la même façon qu'une *CDI->CDD*. De plus, l'ordre des fenêtres n'est pas pris en compte, mais ici encore les données sont souvent assez structurées par le temps pour que la distance restitue la dynamique temporelle. L'avantage de la méthode est de réduire la dimension du tableau à traiter, et peut suffire à rendre compte des grandes oppositions entre trajectoires.

4.4. Sur les comparaisons de méthodes de classification

Comparer les méthodes par leurs résultats peut être, à priori, tentant mais l'opération ne sera pas concluante. En effet, elle dépend de l'approche, est trop délicate à mettre en œuvre de façon équitable pour les méthodes et au final de peu d'intérêt. Chaque méthode a des paramètres propres (qui sont déduits d'hypothèses) dont on ne peut garantir qu'ils soient optimaux (ni même déduits des mêmes hypothèses) : les coûts pour l'optimal matching, les fenêtres pour l'analyse harmonique. C'est aussi le cas du niveau de regroupement qui est différent selon les méthodes et qui n'est pas toujours celui que l'on exploite (par exemple si ce niveau est trop haut ou trop bas). Il y a également le problème de définir une mesure de comparaison et un niveau de significativité, pour déterminer si deux méthodes sont différentes, qui soit équitable et neutre. Le choix de la mesure est en effet dépendant de l'approche : par exemple l'ACM a la faculté de faire émerger à un niveau très haut les populations atypiques de faible effectif. Un test qui mesurera la capacité à mettre en valeur ces populations à un haut niveau de classification mesurera bien les spécificités de l'ACM mais ne sera pas le plus performant pour l'optimal matching.

A cela s'ajoute la question de l'inférence : si on fait varier l'initialisation, ou l'échantillon, le résultat d'un même algorithme varie et donc la mesure de comparaison aussi. Il faut donc commencer par établir des intervalles de confiance par rapport à chaque méthode pour la mesure de comparaison.

Enfin, si on se place dans le cas très simple où la population a une distribution uniforme, toute méthode donnera un bon résultat et les écarts entre méthodes seront importants. Quel sens dès lors donner à un critère qui définirait une meilleure méthode ?

5. DISCUSSION SUR LA MÉTHODE PRESENTÉE ET EXTENSIONS

Certains aspects de la méthode que nous proposons pour analyser les trajectoires se prêtent à la discussion. Les choix effectués lors du regroupement, de l'utilisation de cartes ou du calcul de la distance méritent d'être commentés.

Tout d'abord, l'objectif de visualisation des résultats amène à optimiser une représentation des données en un nombre de dimensions très réduit (en général 2 pour exploiter les supports visuel habituels, papier ou écran). Il est sûr que tout ajustement bidimensionnel de la structure des données, même logique, a ses limites. Dès l'instant où l'on préfère abandonner la représentation et se focaliser sur le seul ajustement, certains algorithmes d'apprentissage des cartes d'auto-organisation optimisent la structure de la carte (elle est alors déterminée par l'apprentissage mais devient impossible à représenter).

La méthode a été présentée sur un exemple où les calendriers sont calibrés (même nombre de mois) et donc, tous censurés à droite au même moment. Elle peut néanmoins se généraliser au cas de calendriers de longueurs variables dès l'instant où la situation peut être définie. C'est-à-dire dans un cas où la composante *temporelle* des situations peut être ramenée à un repère absolu du temps commun à toutes les trajectoires. La matrice des distances intra-situations peut alors être établie et les trajectoires recodées. Le problème devient alors un problème de données manquantes. Si la position temporelle dans un repère absolu ne peut être établie, par exemple avec i des séquences dont on ne sait pas si elles correspondent au début, au milieu ou à la fin de la période d'insertion, cette méthode ne peut plus, de notre point de vue, être appliquée en l'état car elle n'intègre pas dans la distance un critère qui positionne dans le temps les trajectoires avant de les comparer et surtout avant de définir la distance intra-situations.

La méthode, comme pour l'ACM ou l'analyse harmonique utilise une analyse factorielle qui décompose la trajectoire selon des dimensions principales. Cette opération pondère les variables - les situations dans notre cas - selon leur contribution à l'inertie. Dans le cas des données longitudinales, cela revient à sous-pondérer les périodes de stabilité. Cette propriété favorise les analyses où on veut faire apparaître les dynamiques de transitions. Dans notre application, comme définir la fin la phase d'insertion pose problème, cette propriété évite que la phase de stabilité en fin d'insertion marque trop fortement la typologie.

Une extension possible concerne les éléments pris en compte dans le calcul de la distance entre les situations. Dans la version proposée, la matrice *Univers des transitions probables* est triangulaire car on ne prend pas en compte le passé d'une situation. Il est possible d'ajouter des transitions en provenance du passé vers le présent, sans aucune difficulté.

Une autre piste de réflexion concerne le degré de découpage des états, ici d'emploi et de non-emploi. L'usage consiste en effet, dans l'approche discriminante, à catégoriser les états dans une nomenclature grossière et très structurante, en amont de l'analyse statistique (nous avons utilisé dans notre application, à fins de comparaison, un découpage assez commun dans la littérature). Par exemple, les *emplois aidés* recouvrent plusieurs contrats. Les différencier permet de distinguer les performances de chacun en termes d'insertion. En contrepartie, cela demande de travailler avec des états plus nombreux de fréquences plus faibles. Or la méthode présentée s'accommode très bien des états à faible fréquence comme nous l'avons déjà expliqué. Ces états hybrides peuvent donc être différenciés. Cela aura de plus l'avantage de renforcer la pertinence des voisinages et, par voie de conséquence, la robustesse des résultats.

Pour finir avec les extensions, un travail complémentaire a été réalisé sur l'insertion pour mesurer l'influence de différentes dimensions et notamment des caractéristiques des individus à l'ensemble des centres de

classes (Rousset & Giret, 2007). Les résultats obtenus montraient notamment « toutes choses égales par ailleurs » la forte structuration du niveau de diplôme sur les trajectoires d'insertion, un diplôme élevé protégeant des trajectoires les plus précaires. Généralement, cette protection augmente de manière linéaire avec le niveau de diplôme obtenu.

CONCLUSION

De nombreux travaux d'analyse ayant été réalisés sur la structure longitudinale du processus d'insertion des jeunes sur le marché du travail, une grande expérience a été accumulée sur l'intérêt et les limites liés aux différentes approches. Notre objectif a été de tenir compte de cette expérience pour élaborer une méthode susceptible de résumer puis d'expliquer l'hétérogénéité des itinéraires professionnels dans les premières années de vie active. Les données utilisées issues de l'enquête Génération 98 du Céreq nous permettaient d'observer les calendriers professionnels mensuels de 16 000 jeunes sur les sept ans qui ont suivi la fin de leurs études. En utilisant d'une part une distance qui prend en compte certaines spécificités du processus d'insertion à partir des transitions et d'autre part, les cartes d'auto-organisation pour regrouper les trajectoires sur un grand nombre de classes, la méthode retenue nous a permis de mettre en évidence la diversité des parcours professionnels. On obtient ainsi un panorama détaillé mettant en évidence, entre autres, la forte segmentation entre parcours d'accès direct aux emplois stables et parcours dominés par des emplois plus précaires, ainsi que la dynamique temporelle de ce processus et ses différentes étapes.

Par ailleurs, la confrontation de notre méthode à d'autres en usage en illustre les propriétés singulières : la prise en compte de la dynamique dans le temps, l'intérêt de travailler sur un grand nombre de classes et de maîtriser des niveaux fins de regroupement pour comprendre des mécanismes longitudinaux importants pour l'étude de l'insertion. Pour finir cette méthode se généralise, au delà de notre application, au cadre de l'analyse de trajectoires à partir des calendriers.

Remerciements

Les auteurs remercient Ludovic Lebart, directeur de recherche au CNRS, pour ses conseils et ses encouragements.

RÉFÉRENCES

ABBOTT, A. & HRYCAK, A. (1990). Measuring resemblance in sequence data: an optimal matching analysis of musicians' career. *American Journal of Sociology*, 96, 1, 144-185.

BENZECRI J.P., 1973, L'Analyse des Données : TIIB n°2 : représentation Euclidienne d'un Ensemble fini de masses et de distances, Paris, Dunod, p. 65-95.

COTTRELL M., IBBOU S., LETREMY P., ROUSSET P., (2003), « Cartes auto-organisées pour l'analyse exploratoire des données et la visualisation », *Journal de la société française de statistique*, 144 (4), p. 67-106.

COTTRELL, M., GIRARD, B. & ROUSSET, P., (1998), Forecasting of curves using a Kohonen classification, *Journal of Forecasting*, 17, 429-439.

D. DELAUNAY et LELIEVRE E., (2006) –« Examen topographique des transitions biographiques complexes à l'aide des cartes de Kohonen », in *États flous et trajectoires complexes: observation, modélisation, interprétation*, GRAB - Lelièvre & Antoine (Eds). Méthodes et Savoirs n°5, INED/CEPED, Paris, p.219-238.

DEVILLE J.C., (1977), « Analyse harmonique du calendrier de construction des familles en France », *Population*, 32(1), P. 17-63.

ESCOFIER-CORDIER B., (2003), « Analyse factorielle et distances répondant au principe d'équivalence distributionelle. », Analyse des correspondances, Presses Universitaires de Rennes, pp.55-64.

ESPINASSE J.M., (1994), « Enquête de cheminement, chronogrammes et classification automatique », in Ourtau M., Werquin P. (eds.) L'analyse longitudinale du marché du travail, Documents Séminaire Cereq, 99, p. 193-218.

FENELON J.P., GRELET Y., HOUZEL Y., (1997), « Modéliser l'insertion », *Formation-Emploi* , 60, p. 37-48.

FORT J.C., (2006), «SOM's mathematics », Neural Networks, 19, p. 812-816.

GABADINHO A, RITSCHARD G., STUDER M., MÜLLER N.S., (2008), Mining Sequence Data in R with TraMineR : a user's guide, http://mephisto.unige.ch/traminer

GIRET J.F., ROUSSET P., (2005), « Une typologie des débuts de carrière professionnelle en France », 12 th International Meeting of Connectionist Approaches in Economics and Management - ASCEG 2005-, Aix en Provence, novembre.

GRELET Y., 2002, « Des typologies de parcours, méthodes et usages », Notes de travail du Céreq, 20, 47 p.

HALPIN B., CHAN T. W., (1998), « Class careers as sequences : An optimal matching analysis of work-life histories », *European Sociological Review*, 14(2), p. 111-130.

HALPIN B., (2008), « sequence analysis of lifecourse data », Labor Market Changes and Social Exclusion in the Life Course', October 23-25, 2008, Oslo, Norway

KOHONEN T., (2001), Self-Organizing Maps. 3.ed, *Springer Series in Information Sciences*, 30, Springer Verlag, Berlin.

LEBART L., PIRON M., MORINEAU M., (2006), *Statistique exploratoire multidimensionnelle*, Paris, Dunod.

MASSONI S., OLTEANU M., ROUSSET P., (2009), "Career-Path Analysis Using Optimal Matching and Self-Organizing Maps", Advances in Self-Organizing Maps, José C. Principe, Risto Miikkulainen (Ed.) p. 154 pages-162 pages

OJA E. ET KASKI S., (1999), *Kohonen Maps,* Elsevier, Amsterdam.

ROBETTE N., THIBAULT, N., (2008), « Analyse harmonique qualitative ou méthode d'appariement optimal? Une analyse exploratoire de trajectoires professionnelles », *Population*, 63(4), 621-646.

ROUSSET P., GIRET J.F., (2007), « Classifying qualitative time series with SOM: the typology of career paths in France », in Sandoval F., Prieto A., Cabestani J., Grana M. (ed.), *Computation and Ambiant Intelligent, Iwann 2007 proceeding, Lecture Note in Computer Science*, Berlin, Springer, p. 757-764

ROUSSET P., GUINOT C., (2002), « Visualisation des distances entre les classes de la carte de Kohonen pour le développement d'un outil d'analyse et de représentation des données », *Revue de Statistique Appliquée*, 50 (1), p. 35-47.

SCHERER S., (2004), « Stepping-Stones or Traps?: The Consequences of Labour Market Entry Positions for the future career in Germany, Italy and Great Britain », *Work Employment Society*, 18, p. 369-394.

TORGERSON W.S., (1958), *Theory and methods of scaling*, Wiley, New-York.

WU L., (2000), Some comments on "Sequence analysis and optimal matching methods in sociology, review and prospect", *Sociological methods and research*, vol. 29(1), p.41-64

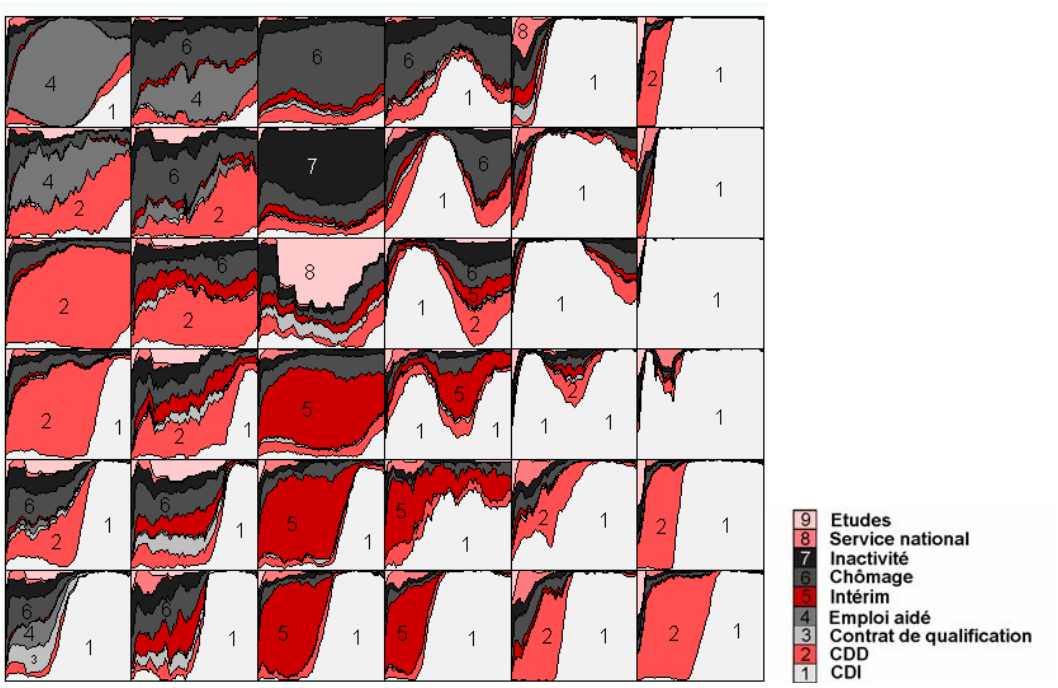
ANNEXE 1 : UNE GRILLE 6X6

La figure 1 présente le résultat de l'application des cartes d'auto-organisation à l'aide d'une grille 6x6 (voir section 3.2). On retrouve une grande cohérence avec la grille 10x10 (section 3.3.1) dans la structure de voisinage des classes. Les contrats courts – *CDD* et *intérim* – sont rapprochés (unités *D*6 et *E*6) dans leur rôle de marchepied vers le *CDI* (contrats précaires en début de période et stables en fin). De même, le *chômage* de longue durée (unité *C*1) et l'*inactivité* (*D*1) sont voisins. Les vagues d'accès au *CDI* via l'*intérim* ou le *CDD* sont détaillés et rapprochés (région Sud-Est de la carte).

Carte 2

Typologie de l'insertion des jeunes sortis de la formation initiale en 1998 sur 7 ans

Classification obtenue à l'aide d'une cartes d'auto-organisation 10x10



ANNEXE 2 : APPLICATION DE DIFFÉRENTES MÉTHODES

On présente ici les applications des différentes méthodes pour définir une typologie des parcours d'insertion au niveau 8. Ces applications sont présentées en complément de la section 4 et sont à relativiser en raison des réserves développées en section 4.4.

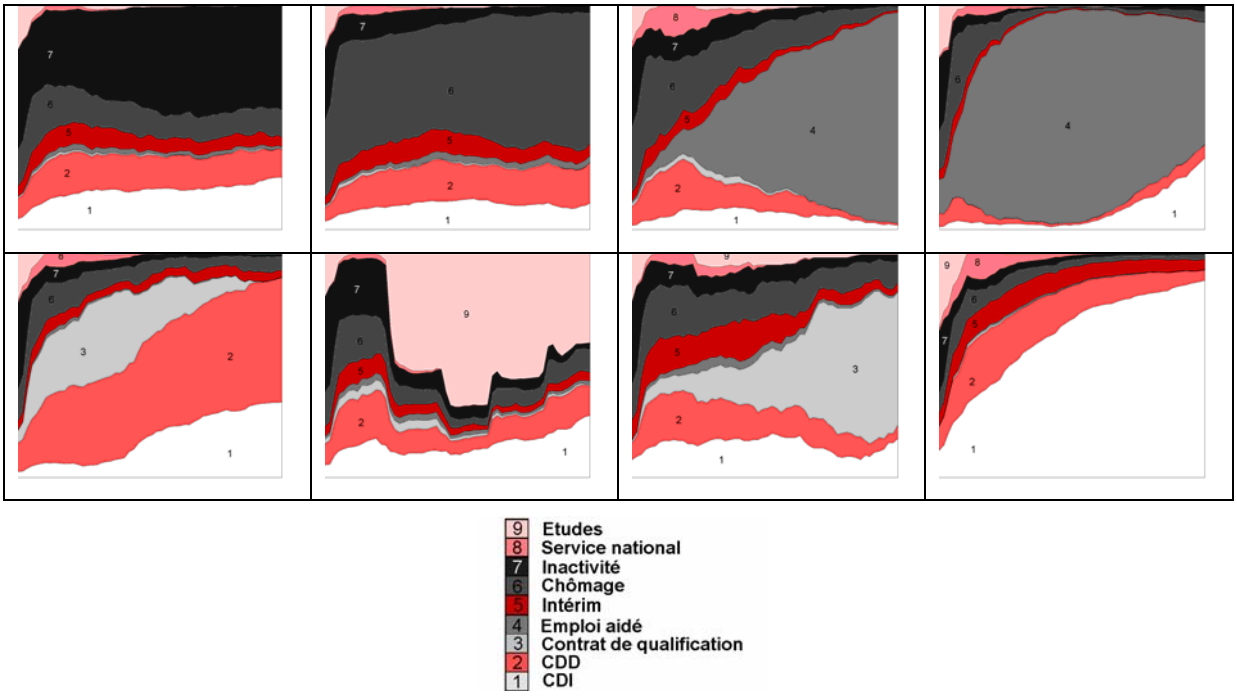
Annexe 2.1 L'Analyse en composantes multiples

Conformément aux caractéristiques théoriques de l'ACM (section 4.1), les classes de faible effectif ont été revalorisées au point d'apparaître à un niveau très haut d'agrégation. Ainsi, les reprises d'études et les contrats de qualification en fin de trajectoire constituent deux classes à faible effectif alors qu'une seule classe décrit l'accès au *CDI*. Cette caractéristique s'amplifie en descendant dans le niveau d'agrégation, des classes à trop faible effectif apparaissant fréquemment. Dans le cas où les classes à faible effectif ne sont pas toutes prioritaires du point de vue de l'étude, il est alors nécessaire de sélectionner dans l'arborescence les branches les plus significatives ou les plus intéressantes par un outil statistique complémentaire ou de façon arbitraire. Pour un regroupement à l'aide des cartes, on pourra se référer à Delaunay et Lelièvre (2006).

Carte 1

Typologie de l'insertion des jeunes sortis de la formation initiale en 1998 sur 7 ans

Partition en 8 classes à partir du couple de l'analyse en composantes multiples



Annexe 2.2 L'optimal matching

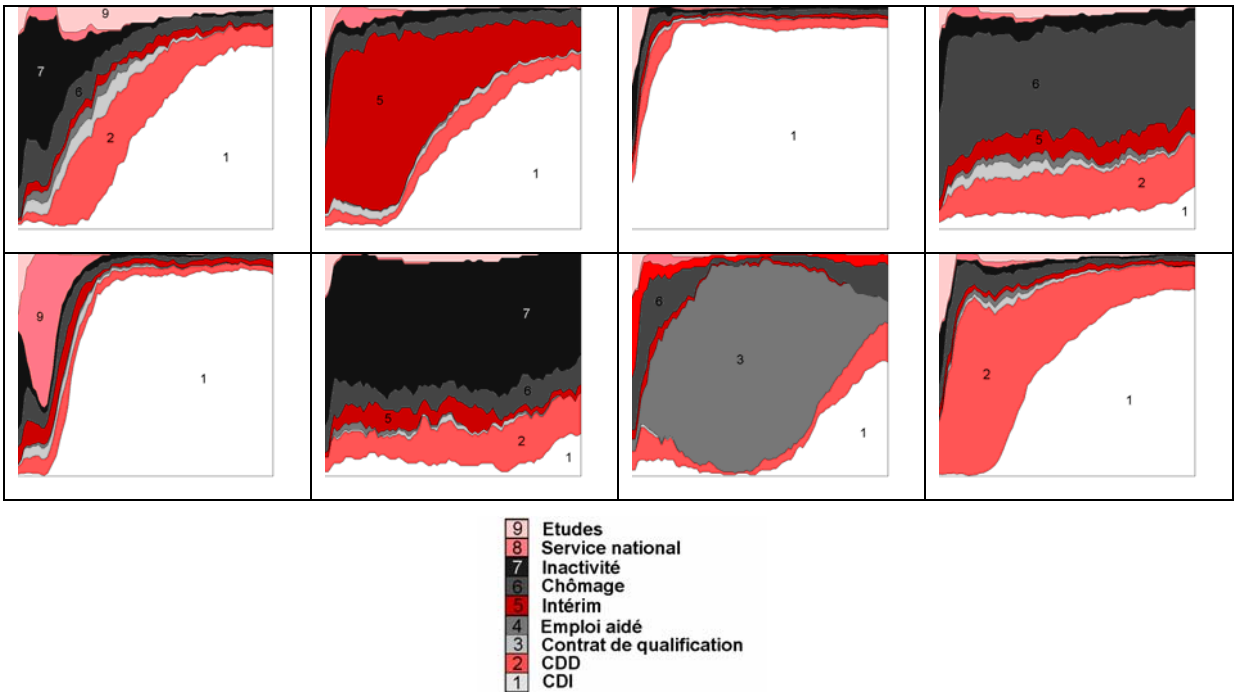
Concernant la partie théorique de l'optimal matching, on pourra revenir sur la section 4.2. Du côté de son application, L'OMA laissant libre le choix des coûts, et donc de la distance entre les trajectoires, l'investigation est très large. Il est difficile d'arbitrer ces coûts dans le cadre de notre comparaison, ce choix ayant évidemment une influence. Dans la littérature, de nombreuses applications proposent de ne considérer que des coûts de 1 ou 0. De même dans Massoni et al (2009) où le package TraManeR (Gabadinho et al, 2008) a été utilisé, la matrice des coûts ne différencie pas nettement les états. Ainsi, par souci d'équité dans le cadre de notre comparaison, nous avons défini des coûts en reprenant les principes de notre méthode même si cela implique des contradictions que nous ne développerons pas. Néanmoins, ces coûts sont calculés à partir des transitions, plus conforme au cadre euclidien (cf. remarque en section 4.2). De plus, une mesure des coûts à partir des transitions s'associe mal avec une transformation ultérieure de ces mêmes trajectoires par l'insertion ou la suppression. Le coût de substitution est calculé de la manière suivante : pour chaque état, on mesure les probabilités de transition à partir de cet état à 1, 2, …, et 12 pas, soit un vecteur à 12 composantes. Le coût de substitution entre deux états est la distance du χ^2 entre les vecteurs de transition respectifs. Nous nous plaçons dans le cas où les coûts suppression/insertion n'interviennent pas. En raison des capacités de nos outils informatiques, nous avons dû réduire la base de données à environ 12270 individus. Enfin, comme nous évoluons dans un cadre discret, la méthode de classification est PAM (patitionning aound medioids).

Au résultat, on observe que le poids du *CDI* et la stabilité ont pesé grandement dans la classification. La stabilité prend plus d'importance si le temps est équipondéré et le *CDI* en fin de période pesant pour 60% de l'effectif n'est pas compensé par le fait que les états de précarité se sont rapprochés et que les contrats à durée déterminée (*intérim* et *CDD*) se sont rapprochés d'eux. Pour un regroupement à l'aide des cartes, on pourra se référer à Massoni et al (2009).

Carte 2

Typologie de l'insertion des jeunes sortis de la formation initiale en 1998 sur 7 ans

Partition en 8 classes à partir de la méthode de l'optimal matching

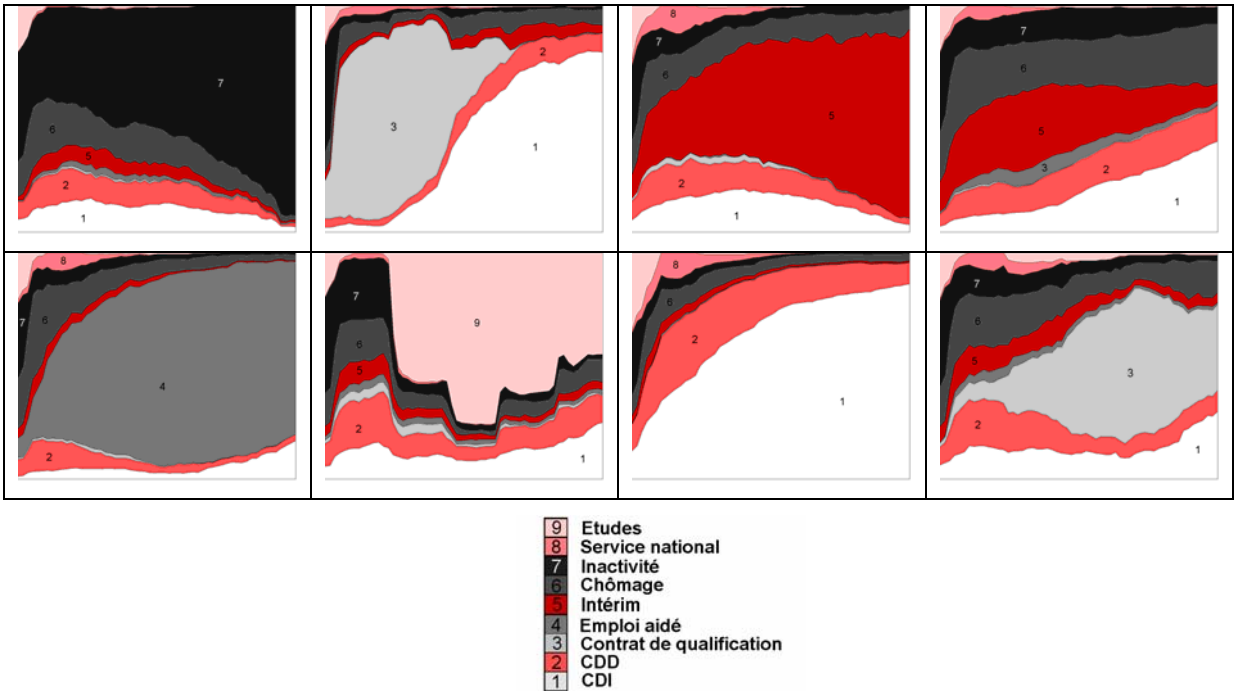


Annexe 2.3 L'analyse harmonique

Comme présenté en section 4.3, nous avons recodé les trajectoires en introduisant des fenêtres de 12 mois. L'analogie de cette méthode avec l'ACM se retrouve dans les résultats : on retrouve les mêmes classes de faibles effectifs à ce haut niveau d'agrégation (*retour aux études* et *contrats de qualification en fin de parcours*) et la classe très nombreuse de l'accès rapide au *CDI* n'est pas d'avantage détaillée. Le découpage en fenêtres du temps ne se retrouve pas dans les structures temporelles des classes. La seule exception (hormis la classe d'accès au *CDI* via *un contrat de qualification*) pourrait être la non présence d'une classe *chômage.* La présence simultanée dans une fenêtre du *chômage* avec les autres états semble avoir été suffisamment fréquente pour avoir été intégrée dans l'analyse et s'être traduite par une dilution du chômage dans les différentes classes.

Carte 3

Typologie de l'insertion des jeunes sortis de la formation initiale en 1998 sur 7 ans. Partition en 8 classes à partir de la méthode de l'analyse harmonique.



ISSN : 1776-3177
Marseille, 2011.