MIGRANTS-MIGRATIONS : UN VIEUX THÈME REVISITÉ

Jean-François ROYER (*)

(*) CREST-INSEE

Introduction

Décrire les évolutions de la mobilité résidentielle sur le territoire d'un pays est un objectif important de la statistique régionale. Mais les migrations intérieures sont des évènements démographiques plus difficiles à manipuler statistiquement que les naissances et les décès. Contrairement à ces derniers, elles ont une définition contingente, qui nécessite le choix d'une convention (souvent un zonage) ; ce ne sont pas des évènements inéluctables ; et ce sont des évènements répétables pour un même individu.

En fait, les outils d'observation démographique fournissent rarement des données directement sur les migrations[1] ; le plus souvent, ils fournissent des données sur les individus d'une population, classés selon qu'ils ont ou non changé de résidence pendant une période, donc des statistiques sur des migrants. Or le nombre de migrations et le nombre de migrants ne sont pas égaux : ils diffèrent d'autant plus que la période d'observation est plus longue. Pour établir des indicateurs standardisés de migration, permettant les comparaisons dans le temps et dans l'espace, il faut traiter cette double question du rapport migrants-migrations et de la longueur de la période d'observation.

En France, Daniel Courgeau s'est attaché à le faire dès la fin des années 1960, et ses travaux ont été repris plusieurs fois depuis ([1] à [5]). Cette question est aussi abordée dans la littérature de langue anglaise, notamment pour comparer les données des recensements avec celles d'enquêtes auprès des ménages ([8]). Dans les deux cas, les auteurs mettent en avant certaines régularités observées du rapport migrants-migrations, et proposent de profiter de ces régularités pour permettre des comparaisons lorsque les périodes d'observation sont de longueurs différentes.

Ces travaux ont été menés à partir de sources qui présentaient de sérieuses limitations : ou bien l'information rétrospective sur les migrations était riche, mais les échantillons étaient très réduits, ou bien les échantillons étaient importants, mais l'information sur les migrations très réduite. Dans cet article, j'utilise les données d'un panel qui n'est pas sans défaut, mais qui est à la fois riche en information sur les parcours résidentiels et de taille importante, le panel tiré des déclarations annuelles de salaires DADS. Ces données permettent de décrire finement, voire de modéliser, les éléments de la relation migrants-migrations. Cela permet d'apprécier la validité des hypothèses des modèles antérieurement proposés pour rendre compte synthétiquement de cette relation, et pour corriger les données brutes en vue des comparaisons.

La section 1 met en place le cadre et les notations, et les illustre en s'appuyant sur le modèle la plus simple possible. La section 2 présente un exemple de statistiques descriptives tirées du panel DADS. La section 3 examine le « modèle de Courgeau » au vu de ces résultats. La section 4 présente un modèle de comportement individuel simple, et son estimation. La section 5 examine si on peut traiter le problème de changement de longueur de période, et la section 6 résume les conclusions.

[1] Je laisse donc de côté les comptages de franchissements, les enquêtes aux frontières, etc.

Section 1 : cadre et notations

Dans tout ce qui suit, il sera question d'une population d'effectif P observée pendant une période [0,T]. Dans la pratique, les échantillons seront donc « cylindrés » : c'est ce qui se passe d'ailleurs nécessairement dès que l'on utilise une question sur la résidence antérieure dans une enquête ou un recensement.
On supposera choisie une définition de l'évènement « migration » (sous-entendu « intérieure »). Ce pourra être, au choix : changement de logement, de commune, de département, de zone d'emploi, de région, de ZEAT. De façon générale, on suppose que les individus peuvent se déplacer dans un ensemble de Z « localisations » ; la localisation initiale (à l'instant 0) et la localisation finale (à l'instant T) joueront des rôles particuliers.
Chaque individu de la population est caractérisé par une chronique de localisations successives pendant la période (temps continu ou temps discrétisé). A partir de ces chroniques, on peut établir les statistiques descriptives suivantes, sur l'ensemble de la population :

- Le nombre de migrations ayant eu lieu dans tout intervalle inclus dans [0,T] : $M[t_1,t_2]$; $M(t)$ désignera le nombre de migrations dans l'intervalle [0,t] ;
- Le nombre d'individus ayant fait au moins une migration entre la date initiale et une date t de l'intervalle [0,T] : $N(t)$; c'est le nombre de migrants ;
- Le nombre d'individus dont la localisation en t diffère de la localisation initiale : $\hat{N}(t)$; c'est le nombre de migrants « apparents ».

Les migrations peuvent être classées en catégories selon leur rang d'occurrence pour un individu donné pendant la période, et selon que leur point de départ est, ou non, la localisation initiale de l'individu. On utilisera les notations suivantes pour les grandeurs relatives à l'intervalle [0,t] :
$PM(t)$: nombre de migrations de rang 1 (égal au nombre des migrants) ;
$SM(t)$: nombre de migrations qui font sortir de la localisation initiale (une migration qui compte dans PM compte aussi dans SM, la réciproque n'est pas forcément vraie) ;
$EM(t)$: nombre de migrations qui font rentrer dans la localisation initiale (retours) ;
$OM(t)$: nombre de migrations dont ni l'origine ni la destination ne sont la localisation initiale (migrations « onward » dans la littérature de langue anglaise).

Par construction :

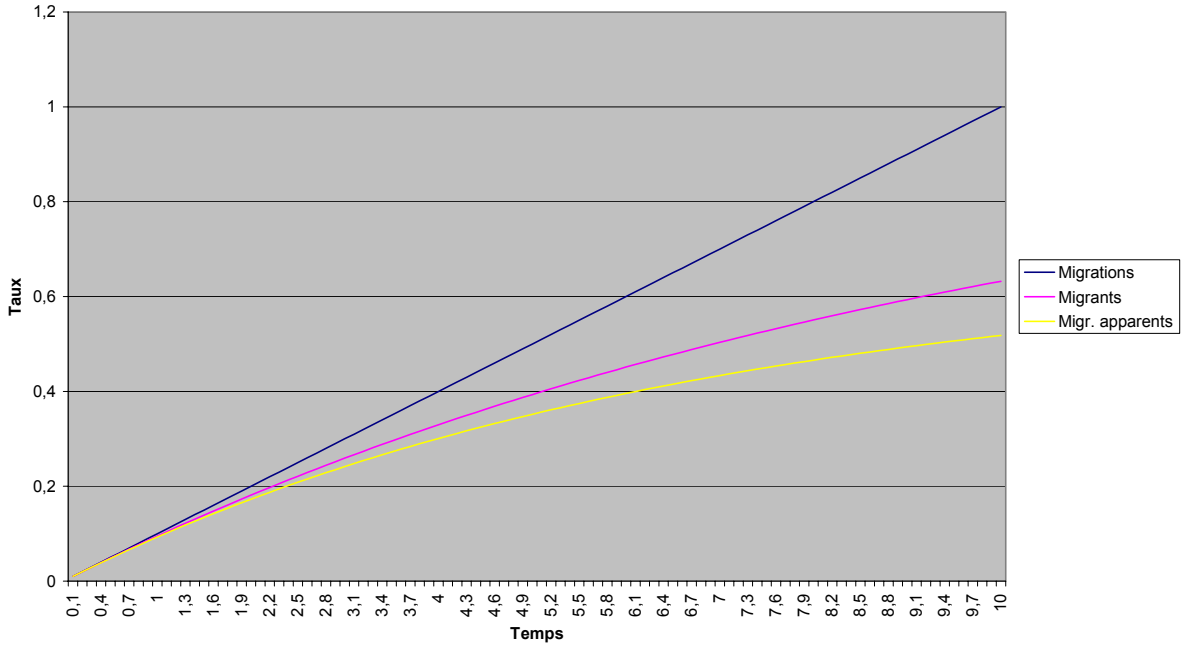
$$N(t) = PM(t)$$
$$M(t) = SM(t) + EM(t) + OM(t)$$
$$\hat{N}(t) = SM(t) - EM(t)$$

Illustration par un modèle ultra-simplifié : pour faire apparaître les relations entre ces différentes grandeurs, on peut supposer que toutes les trajectoires individuelles sont régies par un même processus, et qu'elles sont indépendantes entre elles ; ce processus est une « marche aléatoire » à travers les Z localisations ; à chaque instant un individu a une probabilité instantanée (« hazard ») p de changer de localisation, et, s'il change, les Z-1 autres localisations sont équiprobables.
Sous ces hypothèses, les grandeurs précédentes sont des processus aléatoires fonction de t, et leurs espérances sont (démonstration en annexe 1) :

$$E(M(t)) = Ppt$$
$$E(N(t)) = P(1 - e^{-pt})$$
$$E(\hat{N}(t)) = P\frac{Z-1}{Z}(1 - e^{-\frac{Z}{Z-1}pt})$$

Application : P=1, p=0,1, Z=3

Graphique 7



Ce modèle est irréaliste à maints égards. Deux de ses plus graves défauts sont :

- le fait que dans ce modèle, si t tend vers l'infini le nombre de migrants tend vers P ; à terme, tout le monde sera migrant : or dans les populations réelles il existe des personnes définitivement réfractaires à la migration, « non soumises au risque » ; c'est pourquoi on se tourne vers des modèles du type « migrants-sédentaires » (voir ci-dessous) ;
- le fait que, dans ce modèle, lorsqu'un individu est hors de sa localisation initiale, toutes les localisations autres que celle qu'il occupe se valent pour lui, y compris donc sa localisation initiale ; dans les populations réelles la localisation initiale est choisie préférentiellement.

L'intérêt de ce modèle est de faire apparaître à peu de frais une caractéristique que l'on retrouve dans les observations : la concavité des courbes représentant les nombres de migrants, et de migrants apparents, en fonction du temps.

Section 2 : statistiques descriptives tirées du panel DADS

Les « déclarations annuelles de données sociales – DADS » sont des formulaires administratifs que tous les employeurs sont tenus de fournir une fois par an à la sécurité sociale, aux Impôts et à l'INSEE. Ils sont remplis au niveau de l'établissement (en principe). L'employeur déclare individuellement toutes les personnes qu'il a employées dans l'année écoulée, avec pour chacune l'indication de la période pendant laquelle elle a été employée, et un ensemble d'informations, dont le salaire versé, et aussi l'adresse. Les salariés sont identifiés par leur numéro d'inscription au répertoire des personnes (NIR), ce qui permet de mettre bout à bout les périodes d'emploi d'un même individu au fil du temps[2]. L'INSEE est autorisé à réaliser un tel panel pour un échantillon de salariés au 1/12°[3].
Ce matériau permet de reconstituer la succession des adresses[4] d'un salarié employé continûment pendant une période donnée[5]. S'il existe des « trous » dans la carrière professionnelle d'un individu par rapport au champ des DADS – périodes de chômage, d'inactivité, d'activité non salariée ou salariée dans un secteur non couvert par les DADS comme l'Etat, séjour à l'étranger – on ne connaît pas l'adresse du salarié pendant ces trous : pour établir des données de migrations sur une

[2] Voire simultanément, si une même personne travaille pour plusieurs employeurs au même moment
[3] Depuis 2001 ; auparavant, 1/24°
[4] Adresses connues de l'employeur
[5] A une réserve près concernant les salariés qui changent d'établissement dans une même entreprise en cours d'année : leur changement n'est retracé dans le panel qu'au 1° janvier suivant. Voir ci-dessous.

population qui ne soit pas réduite aux travailleurs continûment employés, on a supposé que l'adresse de la période d'emploi précédente perdurait jusqu'à la veille du premier jour de la période d'emploi suivante. On est obligé de faire une hypothèse « conservatrice » de ce type ; elle sous-estime sûrement quelque peu la mobilité géographique.

Les adresses personnelles disponibles dans les DADS sont codifiées d'après le code officiel géographique[6] : ce matériau permet donc d'étudier la mobilité géographique dans tout zonage supra-communal[7]. Ci-après des résultats sont donnés pour les changements de commune et les changements de région.

Dans les DADS, les périodes d'emploi sont repérées par l'année de déclaration et par une plage de jours dans cette année, les jours étant numérotés conventionnellement de 1 à 360. Les migrations peuvent donc être datées au jour près dans la période étudiée : compte-tenu de la convention indiquée plus haut, une migration est toujours datée du premier jour d'une période d'emploi. Les résultats pourraient être retranscrits par jour ; les graphiques ci-dessous présentent des résultats par mois ou par an.

Les données présentées ci-dessous à titre d'exemple sont relatives aux individus de la génération née en 1970, observée entre 1995 et 2005. Trois conditions doivent être remplies pour qu'un individu appartienne à l'échantillon retenu : il doit avoir une période d'emploi commençant en janvier 1995 ; il doit avoir aussi une période d'emploi commençant en janvier 2005 ; toutes ses périodes d'emploi doivent être localisées en France métropolitaine. Le nombre de trajectoires retenues est de 22 886[8].

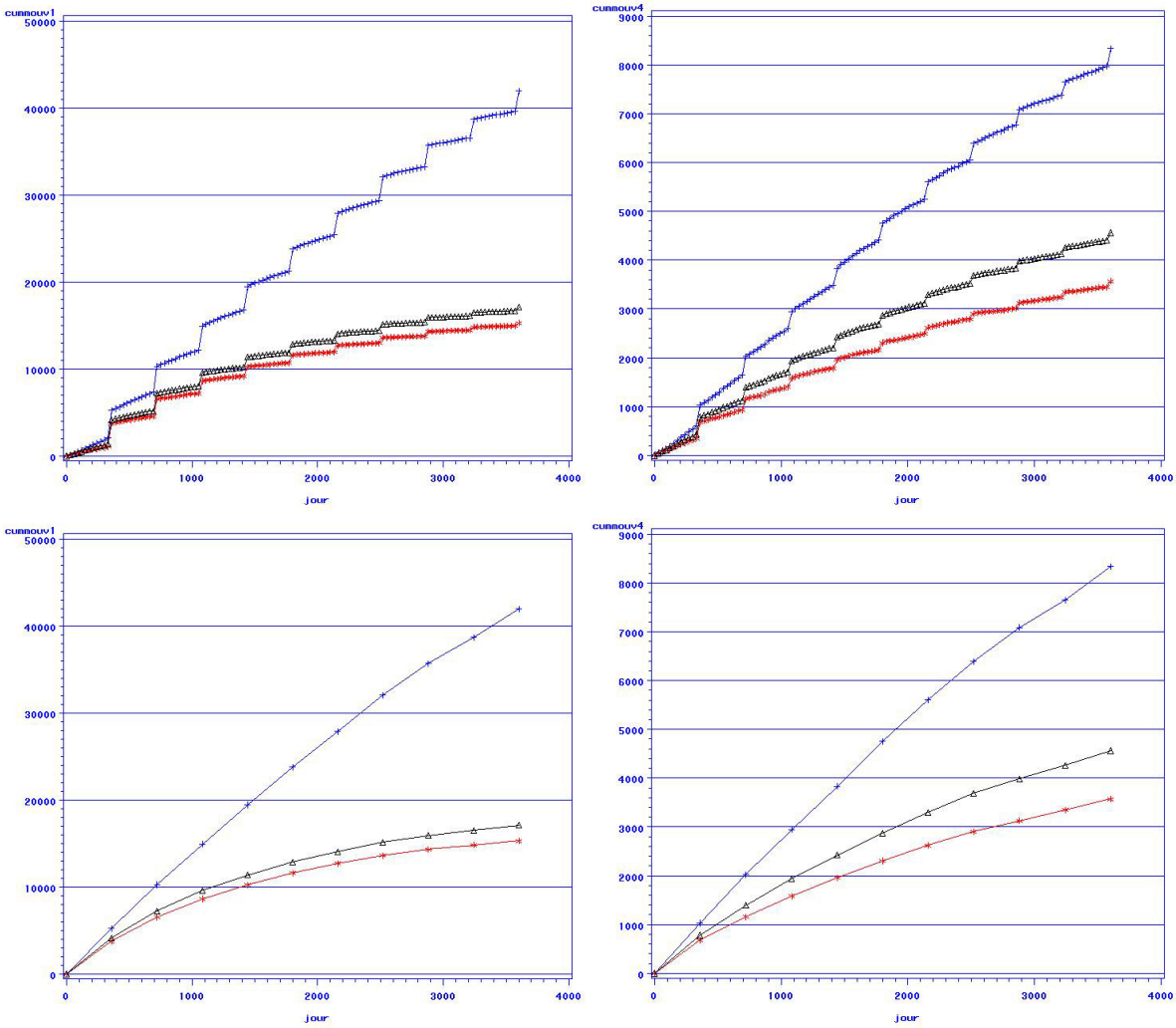
Tableau 1 : Migrations et migrants parmi les salariés du champ DADS de la génération 1970 entre 1995 et 2005

	Changements de commune	Changements de région
Nombre total de trajectoires P	22 886	22 886
Nombre total de migrations $M(10)$	42 065	8 376
Dont : Premières migrations $PM(10)$	17 119	4 577
Migrations « plus loin » $OM(10)$	16 131	1 044
Migrations de retour $EM(10)$	5 308	1 880
Départs d'ordre >1 $SM(10) - PM(10)$	3 507	875
Migrants=premières migrations $N(10)$	17 119 *(75% de P)*	4 577 *(20% de P)*
Migrants apparents au bout des 10 ans $\hat{N}(10)$	15 318	3 572

[6] A compter de l'année 1988

[7] On peut aussi se référer à des distances parcourues, calculées à vol d'oiseau entre les centroïdes des communes

[8] Au recensement de 1999, on a dénombré 531 000 salariés (hors Etat) dans cette génération ; ce qui donnerait 44 000 individus dans un sondage au 1/12°. Dans le panel DADS, on trouve 39 000 individus de cette génération ayant un emploi début 1995 ; sur ces 39 000, seuls 22 886 vérifient aussi les deux autres conditions.



Courbe supérieure : $M(t)$

Courbe intermédiaire : $N(t)$

Courbe du dessous : $\hat{N}(t)$

Les graphiques de gauche sont relatifs aux changements de commune, ceux de droite aux changements de région.

Les graphiques du dessus montrent un point par mois ; ceux du dessous seulement un point par an (mêmes données).

Commentaires :
La convention prise pour constituer le panel DADS (voir note 5) explique les « sauts annuels » des graphiques mensuels : des migrations intervenues en cours d'année sont artificiellement reportées au 1° janvier suivant.
Le nombre annuel de migrations est presque constant les premières années, il décroît ensuite au fur et à mesure du vieillissement de la génération.
La courbe du nombre de migrants ressemble à la courbe théorique présentée dans le paragraphe précédent, mais son asymptote ne correspond pas au nombre total d'individus : il y a des « sédentaires » définitifs.
La courbe du nombre des migrants apparents s'écarte plus de celle des migrants dans le cas des migrations interrégionales que dans le cas des migrations intercommunales : il y a relativement plus de retours dans le premier cas.

Compte-tenu de la taille du panel, ce genre de résultat peut être établi pour différentes générations, différentes périodes, différentes sous-populations (par sexe, par région de naissance ou de résidence initiale, etc.). Cette source, si elle est gérée de manière stable, peut apporter des éléments au suivi annuel de la mobilité géographique. On peut s'en servir aussi pour tester les modèles disponibles sur le sujet.

Section 3 : le « modèle de Courgeau »

Dans une série d'articles publiés dans Population, revue de l'INED, à partir de 1973, Daniel Courgeau a proposé une relation connue depuis, en France, sous le nom de « modèle de Courgeau ».
Pour une part, il s'agit d'un modèle de comportement individuel : Courgeau suppose que, dès lors qu'un individu a déjà effectué une migration, cet individu a une probabilité K d'en effectuer une nouvelle, et que cet évènement survient aléatoirement avec une probabilité instantanée constante p. K ne dépend que du découpage géographique utilisé ; p est une constante valable quel que soit le rang de la migration et le découpage géographique.
Ceci ne suffit pas à modéliser entièrement le phénomène, puisque la première migration reste non spécifiée.
La méthode de Courgeau consiste à se contenter de ces hypothèses pour établir une relation entre le taux de migrants observés pendant une période [0,t], soit selon nos notations $N(t)/P$, et une grandeur (notée m) qu'il appelle « quotient instantané de migration » et qui correspond dans nos notations à la dérivée de la fonction $M(t)/P$. La formule est établie en supposant ce « quotient » constant sur la période, mais peut se généraliser sans difficulté au cas où il varie selon une fonction connue du temps. Le nombre de migrations se déduit directement de ce quotient : $M(t) = Pmt$ si m est fixe, $M(t) = P\int_0^t m(\tau)d\tau$ si m varie au cours du temps.

Dans ce cadre, Courgeau montre (démonstration rappelée en annexe 2, pour le cas de m fixe) que :

$$N(t)/P = m[(1-K)t + (K/p)(1-e^{-pt})] = mt(1-K(1-\frac{1-e^{-pt}}{pt}))$$

Sachant que $m = \frac{dM(t)}{Pdt}$, et donc que si m est constant $m = \frac{M(t)}{Pt}$, on trouve :

$$N(t) = M(t)(1-K(1-\frac{1-e^{-pt}}{pt}))$$

(remarque : si K=1, on retrouve la formule du modèle de la section 1)

Prise en compte des retours : Courgeau se contente d'affirmer qu'une proportion l des migrations de rang supérieur à un correspond à des retours. Il ne peut pas s'agir d'un paramètre de comportement (ne serait-ce que parce qu'une migration suivant un retour ne peut pas être elle-même un retour !), et l'interprétation de cette grandeur est assez obscure. Après cette introduction, sa formule devient :

$$\hat{N}(t) = M(t)(1 - K(1+l)(1 - \frac{1-e^{-pt}}{pt}))$$

Commentaire : Le modèle de Courgeau est un être complexe ! S'il s'agissait d'un modèle de comportement « ordinaire », il spécifierait toutes les transitions, y compris les primo-migrations, et le nombre de migrations ou le nombre de migrants seraient des processus aléatoires résultant de cette spécification, dont on pourrait calculer les moments en fonction des paramètres.
En réalité seuls les mouvements de rang supérieur à un sont spécifiés à l'aide de paramètres, en fonction des mouvements de rang un ; et c'est une équation de nature comptable qui relie les statistiques globales observables entre elles en présence de ces paramètres (voir la démonstration en annexe). Les statistiques observables ne sont donc pas déduites du modèle, mais seulement « reliées » par le modèle.
Les mouvements de rang un sont « implicitement spécifiés » si on fixe la forme de $M(t)$. Par exemple, si on suppose que les migrations sont proportionnelles au temps, de la forme mPt, on peut en déduire la variation du risque de primo-migration avec le temps. Le nombre de primo-migrations entre t et t+dt s'obtient en différenciant le nombre de migrants ; la population soumise au risque est la population totale moins le nombre de migrants en t.
On trouve :

$$risquedeprimo - migration = m[1 - K - Ke^{-pt}]/(1 - m[(1-K)t + (K/p)(1-e^{-pt})])$$

En tant que modèle de comportement, ce n'est pas très facile à justifier !
De ce caractère très particulier du « modèle », il résulte des particularités tant pour l'estimation des paramètres que pour l'utilisation en dehors de la période d'estimation.

Estimation des paramètres : Deux types d'estimation ont été pratiquées :

- estimation à partir d'une seule source, une enquête rétrospective donnant un calendrier des mobilités géographiques : c'est l'estimation que Daniel Courgeau a pratiquée dans ses premiers articles, utilisant des enquêtes par sondage de l'INED des années 60-70 ; les résultats sont des valeurs de K, p, l pour plusieurs types de mobilité.
- estimation à partir de trois sources : une enquête du type précédent, une série d'enquêtes emploi (question rétrospective sur un an), et un recensement (question rétrospective sur 7 ou 9 ans) : c'est l'estimation que Daniel Courgeau a pratiquée dans les années 80 sur la période 1975-1982, et que Franck L'Hospital a renouvelée en 2001 sur la période 1990-1999. L'enquête avec calendrier est utilisée uniquement pour établir une valeur de p, mais pas de K (à cause de la faiblesse de l'échantillon) ni de l (parce que l'information sur les retours manque) ; le recensement fournit les migrants apparents $\hat{N}$ pour la période pluriannuelle et les enquêtes emploi fournissent les migrations M pour chacune des années qui composent cette période[9] ; on en déduit, utilisant la valeur précédente de p, une valeur de $K(1+l)$.

Utilisation en dehors de la période d'estimation : Pour utiliser le modèle de Courgeau à propos d'une nouvelle période une fois que celui-ci a été étalonné sur une période de base, il ne suffit pas de supposer constants les paramètres de comportement ; il faut supposer qu'on connait un des résultats – migrants ou migrations – pour calculer l'autre. Le plus souvent, ce sont des statistiques de migrants apparents sur une période longue, intercensitaire, qui sont ainsi « converties » en statistiques de migrations[10]. L'hypothèse est la suivante : s'il y a eu variation des paramètres de mobilité entre la période de base et la période nouvelle, ces variations n'affectent pas les paramètres des mobilités de rang 2 et plus ; donc on va pouvoir en mettre les effets en évidence en utilisant la relation de Courgeau pour ramener les deux observations – période de base et nouvelle période – à une même unité de temps ; les différences statistiques que l'on constatera alors seront significatives de modifications des comportements de mobilité. Implicitement, ce raisonnement suppose que les variations des comportements de mobilité n'interviennent que sur les primo-mobilités.
Mutatis mutandis, c'est le même type de raisonnement qui est sous-jacent lorsqu'on utilise la relation de Courgeau au sein d'une même période, sur une sous-population de la population pour laquelle les

[9] En toute rigueur, une enquête emploi ne fournit pas les migrations sur un an, mais les migrants apparents sur un an ; cette durée étant jugée courte, les travaux cités ici ont assimilé les deux.
[10] Ou de migrants sur un an, par assimilation : cf. note 7

paramètres K, p, l ont été estimés : on suppose que les disparités migratoires entre sous-populations n'interviennent que sur les primo-mobilités.

Section 4 : un modèle de comportement du type « mobiles-sédentaires »

Il est possible de spécifier un modèle de comportement encore rudimentaire, mais qui fasse une place aux deux idées que le modèle ultra-simplifié de la section 1 ne prenait pas en compte : l'existence de « sédentaires » non soumis au « risque de mobilité », l'existence d'un attachement particulier à la localisation de départ. Ce modèle est inspiré des idées de Courgeau pour décrire les migrations de rang 2 et plus, mais il spécifie aussi la première migration, ce qui lui permet d'être plus simple conceptuellement.

La première hypothèse est qu'un individu i de la population appartient a priori (et donc de façon exogène) à une et une seule des catégories suivantes : sédentaire, 1-mobile, 2-mobile, etc. Un individu sédentaire n'est jamais soumis au risque de mobilité ; un individu 1-mobile est soumis au risque de mobilité jusqu'à ce qu'il ait effectué sa première migration, après quoi il ne l'est plus ; un individu n-mobile est soumis au risque de mobilité jusqu'à ce qu'il ait effectué sa n-ième migration, après quoi il ne l'est plus.

Deux paramètres décrivent la répartition des individus entre ces catégories : K_1 la proportion de sédentaires, K_2 la proportion de ceux qui s'arrêtent au niveau n parmi ceux qui sont au moins n-mobiles, pour tout $n \geq 1$. On en déduit que la proportion des « n-mobiles » est :

$(1-K_1)(1-K_2)^{n-1}K_2$.

La seconde hypothèse décrit les probabilités de transition entre états. Les états possibles pour un individu i sont décrits par deux variables : le nombre de migrations qu'il a déjà effectuées depuis l'instant initial, d'une part, et le fait qu'il se trouve, ou non, dans sa localisation initiale d'autre part. Pour un individu qui est encore soumis au risque de migrer, ce risque sera supposé constant et donc indépendant du temps écoulé depuis l'instant initial ou depuis la migration précédente. Quatre paramètres suffisent pour spécifier toutes les transitions possibles :

- pour un premier mouvement : p_1
- pour un mouvement de rang >1 partant d'une localisation qui n'est pas la localisation initiale : p_3 si ce mouvement est « vers l'avant », r si c'est un retour
- pour un mouvement de rang >1 partant de la localisation initiale : p_2

Le modèle est spécifié en temps discret : l'intervalle [0,T] correspond à T transitions successives. Les individus sont supposés indépendants.

La vraisemblance d'une trajectoire peut alors se calculer facilement (voir annexe 3). On peut estimer les 6 paramètres par la méthode du maximum de vraisemblance.

Estimation : à titre d'exemple, voici les résultats obtenus en appliquant ce modèle aux données décrites plus haut (section 2). Il s'agit donc d'un échantillon de 22 886 individus de la génération 1970 observés entre janvier 1995 (localisation initiale) et janvier 2005.

Tableau 2 : Estimation d'un modèle « mobiles-sédentaires » sur les salariés du champ DADS de la génération 1970 entre 1995 et 2005

Paramètres	Changements de communes	Changements de région
$100K_1$	10.7 (0.6)	65.1 (1.5)
$100K_2$	16.0 (0.4)	43.0 (0.7)
$10000p_1$	5.1 (0.08)	2.4 (0.2)
$10000p_2$	14.4 (0.32)	19.3 (0.8)
$10000p_3$	6.4 (0.07)	4.3 (0.2)
$10000r$	2.1 (0.04)	7.7 (0.3)

Estimations des paramètres avec entre parenthèses l'erreur standard

L'unité de temps pour les paramètres p_1, p_2, p_3, r est le jour : multiplier par 360 si on veut comparer à des estimations issues d'études sur base annuelle.

Commentaires :
Les estimations ne sont pas invraisemblables. Les pourcentages de sédentaires – 11% et 65% - sont nettement inférieurs aux pourcentages de « non-migrants au bout de dix ans » du tableau 1 – 25% et 80%.

La part des retours dans les mouvements qui pourraient être des retours est $\frac{r}{r+p_3}$: elle vaut 25% pour les changements de commune, 64% pour les changements de région.
Tous les paramètres changent significativement quand on change de niveau géographique.

Section 5 : le problème du changement de longueur de période

Souvent, on veut comparer des statistiques de migrants relatives à des périodes disjointes de longueurs différentes : par exemple, les intervalles intercensitaires entre 1954 et 1999. Ou encore, on veut comparer des statistiques de migrants sur une même période provenant de sources dont l'intervalle d'observation rétrospective diffère : par exemple, période 2001-2006 vue à travers le recensement 2006 (question rétrospective sur 5 ans) et les enquêtes emploi 2002-2006 (questions rétrospectives sur 1 an).
On se trouve dans la situation où on ne connaît que la statistique des « migrants apparents » au sens des graphiques ci-dessus, pour des intervalles de longueur différente. Sur ce genre de statistique, la division par la longueur de la période n'est pas la bonne solution, du fait de la concavité des courbes.

Référence américaine :
Pour utiliser les données de l'American community survey, des géographes américains se sont posé une question de ce genre (Rogers, Raymer, Newbold 2002). Leur problème est de passer de 5 ans (période utilisée dans les recensements traditionnels) à 1 an (période utilisée dans l'ACS). Ils profitent du fait que les résultats du Census sur 5 ans peuvent être rapprochés des résultats sur 1 an du « Current population survey » CPS. Ils proposent deux indices très simples :

- soit $5\hat{N}(1)/\hat{N}(5)$, qui représente approximativement un nombre moyen de migrations par migrant apparent ; ils appellent cela « indice de Long-Boertlein » ;
- soit $\hat{N}(5)/\hat{N}(1)$, qui répond à la question : par combien faut-il multiplier le nombre de migrants apparents sur 1 an pour obtenir le nombre de migrants apparents sur 5 ans ? Ils appellent cela « rescaling factor », « facteur de changement d'échelle ».

Avec les données DADS des graphiques ci-dessus, le premier calcul donnerait 1,5 au niveau communal, 1,3 au niveau régional ; le second donnerait 3,3 et 3,8.
Voici un exemple de déclinaison par grande région des États-Unis et du Canada pour le « rescaling factor » :

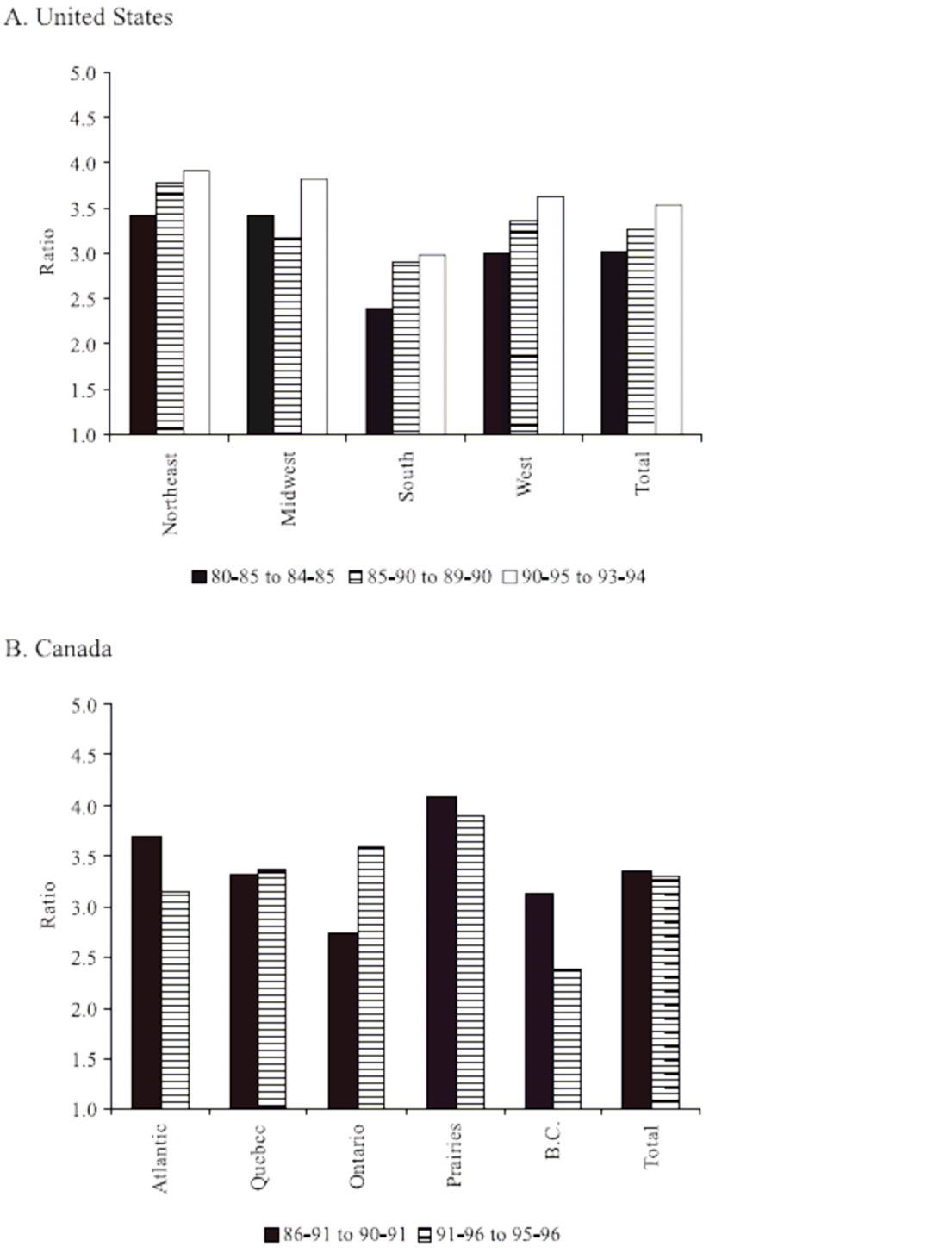


Fig. 1. Observed U.S. and Canadian five-year to one-year ratios of aggregate and regional out-migrant proportions

Source : [8] page 585

On voit que ces ratios varient beaucoup dans le temps et dans l'espace !

Référence française :

La pratique française a été de s'appuyer sur la relation de Courgeau « étalonnée » sur une période antérieure[11]. Le résultat est une estimation de la mobilité totale (approximée par le pourcentage de migrants apparents sur une courte période), reposant sur l'hypothèse que les migrations de rang supérieur à 1 restent régies par des paramètres constants. Cette hypothèse est douteuse, et du coup la comparaison des niveaux de mobilité ainsi obtenus pour des périodes intercensitaires successives est également entachée d'un doute : cf. Royer 2007.

Le paramètre p figurant dans la relation de Courgeau était pris égal à 0,18[12] dans tous les travaux menés de 1973 à 2000, il est passé à 0,26 dans les travaux postérieurs à 2001. En prenant le cas où $K(1+l)=0{,}8$ (ordre de grandeur courant dans ces travaux), ce changement de p à lui tout seul fait passer l'indice $\hat{N}(5)/\hat{N}(1)$ de 4,05 à 3,89 soit une variation de 4%, du même ordre de grandeur que beaucoup des variations des quotients de mobilité totale obtenus par cette méthode entre périodes intercensitaires successives. La modification de $K(1+l)$ a des effets similaires : si ce paramètre passe de 0,77 à 0,59, comme il l'a fait pour le niveau régional en passant des estimations de 1954-1975 à celles de 1975-1990 ([5]), avec une même valeur pour p (0,18), l'indice $\hat{N}(5)/\hat{N}(1)$ varie de 4,17 à 4,50.

Alors, que faire ?

Ce qui précède montre qu'il n'y a pas de méthode miracle pour corriger des différences de longueur des périodes. Si on connait, par d'autres sources, des résultats plus fins par sous-période pour la même population, on n'a pas vraiment besoin de corriger. En cas contraire, on ne peut pas affirmer qu'on connait bien les paramètres de comportement sous-jacents, ni même une partie d'entre eux ; une correction fondée sur des paramètres hypothétiques introduit un « bruit » du même ordre de grandeur que les évolutions qu'on veut étudier.

S'il s'agit seulement de disposer d'ordres de grandeur, des statistiques passées comme celles établies dans l'étude américaine peuvent les fournir, sans qu'il soit besoin de modèle : leur variabilité sera un bon garde-fou. On peut faire ce travail à partir de la confrontation enquêtes emploi-recensements, ou à partir du panel DADS. Mais il faut se garder de penser que de tels « facteurs de changement d'échelle » aient valeur générale : ils dépendent de la région, du groupe d'âge, de la période, etc.

Pour discerner les évolutions ou faire des comparaisons spatiales, on doit utiliser des périodes de longueurs égales. Heureusement, le nouveau recensement, désormais, le permet !

Section 6 : conclusions

Les statisticiens intéressés par les questions locales ont beaucoup plus l'habitude de traiter des enquêtes ponctuelles ou des recensements que des panels. Lorsqu'ils veulent parler de migrations internes, ils n'ont alors qu'un seul type de statistique sous la main : les flux de migrants apparents[13] entre deux dates, la date d'enquête et une date antérieure d'une année ou de quelques années. Une tendance naturelle, à laquelle il faut résister, est de croire que « ce qu'on voit est la totalité de ce qui existe ». Ce n'est pas le cas. Les flux migratoires « onward », les flux de retour, les re-migrations après retour existent, et non seulement ces flux existent mais ils sont numériquement importants dès que la période d'observation rétrospective est un peu longue.

Derrière la question, qui semble technique, de la différence entre statistiques de migrants et statistiques de migrations, il y a le traitement de ces flux considérés comme « secondaires ». Lorsque l'on veut comparer deux sources ayant des périodes d'observation de longueurs différentes, ces flux apparaissent comme des « perturbations » qu'on est tenté de « corriger ». Dans cet article, j'ai voulu d'une part donner une idée de l'importance des différents types de flux, à l'aide d'un panel de salariés,

[11] Voire sur la même période, lorsqu'ont été appliquées aux données d'un recensement les coefficients de Courgeau estimés en utilisant ce même recensement et les enquêtes emploi de la période intercensitaire.

[12] Ici l'unité de temps est l'année

[13] A sens donné à ce mot ici : localisation différente à la date d'observation et à la date antérieure

le panel DADS ; d'autre part examiner si, quand on ne dispose que d'une enquête classique, on peut estimer ces flux « secondaires » pour « corriger » les flux observés.
On ne peut pas. Du moins, pas mécaniquement. Les comportements de deuxième migration (et plus) peuvent être modélisés, et on peut leur appliquer les techniques économétriques connues : mais il est illusoire de penser qu'ils sont beaucoup plus stables et réguliers que les comportements de première migration. Supposer qu'ils sont régis par un petit nombre de paramètres fixes permet d'en donner un bon ordre de grandeur, mais ne peut être considéré que comme une approximation, susceptible d'affecter des comparaisons. En réalité, ces flux méritent d'être étudiés au même titre que les premières migrations, avec prise en compte de co-variables pertinentes : sexe, niveau d'éducation, région d'origine, etc. ; et avec l'idée qu'ils évoluent dans le temps, et pas toujours en phase avec les premières migrations. En particulier, il semble bien qu'on constate un affaiblissement des flux de retour (Royer 2007).
L'âge de l'individu concerné est un facteur essentiel de son comportement migratoire. C'est pourquoi dans toutes les études de comportement, le choix du point de départ de la période d'observation est crucial. On ne peut pas s'attendre aux mêmes résultats si on observe une génération à partir de 25 ans (comme cela a été fait ci-dessus dans les sections 2 et 4), ou à partir de 15 ans. C'est une raison de plus pour laquelle il est irréaliste d'espérer que des coefficients universels fixes permettent de passer du nombre de migrants sur 5 ans au nombre de migrants sur 1 an pour tout groupe démographique, quelle que soit sa composition par âge.
Dans un texte de 2005 ([1]), Bernard Aubry avait déjà attiré l'attention sur ce point : il écrivait en particulier « cela ([l'application généralisée du modèle migrants-migrations aux résultats du recensement]) impliquerait aussi que l'on ait la certitude que les paramètres ne sont pas trop sensibles aux différences de structure (notamment structure par âge dans les zones d'origine et de départ ».
J'espère avoir montré que le panel DADS peut à l'avenir permettre des études de migrations internes très intéressantes, y compris au niveau régional. Sa limitation principale provient de sa source administrative : on ne sait pas ce qu'ont vécu les individus observés en dehors de leurs périodes d'emploi salarié. Des appariements « sécurisés » devraient permettre à l'avenir de compléter cette information : en ce qui concerne les périodes de chômage, ce devrait être pour bientôt. L'enjeu le plus important serait de mieux raccorder les périodes d'études après la fin de la scolarité obligatoire, et les périodes d'emploi.

Bibliographie :

[1] Aubry B., « Les flux résidentiels vus par les recensements et les enquêtes emploi – 1982-1999 » (notamment la deuxième partie : « Examen du modèle migrants-migrations » Note interne INSEE-Alsace 31 janvier 2005
[2] Baccaïni B., Courgeau D., et Desplanques G. « Les migrations intérieures en France de 1982 à 1990. Comparaison avec les périodes antérieures » *Population INED* vol.48 n°6 1990
[3] Baccaïni B. « Enquêtes annuelles de recensement – Résultat de la collecte 2004 – Des changements de région plus fréquents qui bénéficient aux régions du sud et de l'ouest » *Insee Première n°1028* – Juillet 2005
[4] Baccaïni B. « Observations et concepts en matière de migrations internes - Chapitre 122 in Démographie : analyse et synthèse Tome VIII Observation, méthodes auxiliaires, enseignement et recherche » *Editions de l'INED* 2006
[5] Courgeau D. « Migrants et migrations » *Population INED* vol.28 n°1 1973
[6] Courgeau D. et Lelièvre E. « Estimation des migrations internes de la période 1990-1999 et comparaison avec celles des périodes antérieures » *Population INED* vol.59 n°5 2004
[7] Frydman H. « Maximum likelihood estimation in the mover-stayer model » *Journal of the american statistical association* vol.79 n°387 pp.632-638 1984
[8] L'Hospital F. « Les migrations internes en France – Estimation des paramètres du modèle migrants-migrations de Daniel Courgeau » *Note interne INSEE-Rhône-Alpes* Septembre 2001
[9] Rogers A., Raymer J., Newbold K.B. « Reconciling and translating migration data collected over time intervals of differing widths » *The annals of regional science* vo.37 pp.581-601 2003
[10] Royer J.F. « Quatre observations sur la mobilité résidentielle en France métropolitaine », *document de travail du CREST* n°2007-10

Annexe 1 : marche aléatoire à travers Z localisations

On raisonne en temps continu sur $[0,t]$. Soit $X(t)$ le numéro de la localisation dans laquelle se trouve l'individu considéré (on omet son indice) à la date t.
On introduit la notation $p_{ij}(t) = P[X(t) = j / X(0) = i]$; $\vec{p}_i(t)$ désigne le vecteur-ligne des $p_{ij}(t)$, $\vec{p}'_i(t)$ sa dérivée par rapport au temps. Les hypothèses sur les transitions équivalent à :
$\vec{p}'_i(t) = \vec{p}_i(t).Q$ où Q est une matrice (Z,Z) dont les éléments diagonaux valent $-p$ et les éléments hors diagonale valent $\frac{p}{Z-1}$.
Il suffit de résoudre une de ces équations différentielles (tout est symétrique par rapport aux Z localisations) en utilisant les valeurs propres et vecteurs propres de Q :

$$Q = \frac{p}{Z-1}B \text{ avec } B = \begin{pmatrix} -(Z-1) & & \dots & 1 \\ 1 & -(Z-1) & \dots & 1 \\ 1 & 1 & -(Z-1) & 1 \\ 1 & 1 & \dots & -(Z-1) \end{pmatrix}$$

B a Z valeurs propres : 0, et (-Z) avec multiplicité Z-1 . Les vecteurs propres associés sont $\begin{pmatrix} 1 \\ 1 \\ \dots \\ 1 \end{pmatrix}$ et

$\begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}$, etc.

La solution de l'équation permet de calculer $p_{11}(t) = \frac{1}{Z}[1 + (Z-1)e^{-\frac{Z}{Z-1}pt}]$

L'espérance du nombre de migrants apparents est $E(\hat{N}(t)) = P(1 - p_{11}(t))$ d'où la formule du texte.

Annexe 2 : démonstration de Daniel Courgeau

On traite d'une population stationnaire d'effectif P dans laquelle le quotient instantané de migration m est constant. On observe cette population dans l'intervalle de temps [0,t].
Entre les instants θ et $\theta + d\theta$, $Pmd\theta$ individus effectuent une migration. Parmi ceux-ci, $KPmd\theta$ re-feront une migration un jour, K étant une constante. Le temps d'attente de cette nouvelle migration est une variable aléatoire de loi exponentielle avec un paramètre p, autre constante. A un instant τ postérieur à θ, la fraction des $Pmd\theta$ individus qui aura déjà fait cette re-migration est alors $1 - e^{-p(\tau-\theta)}$; c'est en particulier le cas quand $\tau = t$. Il est alors possible de dénombrer les migrations de rang 1 intervenant dans l'intervalle [0,t], sachant que par « rang 1 », on entend « première migration depuis l'instant 0 ». Il suffit pour cela de déduire du nombre total des migrations, qui vaut Pmt, le nombre des re-migrations « induites entre θ et t » par les migrations de l'instant θ, ceci pour toutes les valeurs de θ entre 0 et t.

Donc : migrations de rang 1 = $Pmt - \int_{\theta=0}^{\theta=t} KPmd\theta(1-e^{-p(t-\theta)}) = Pmt - PmpK[t-(1-e^{-pt})/p]$

Et : migrations de rang 1 = $Pm[(1-K)t + \frac{K}{p}(1-e^{-pt})]$

Annexe 3 : vraisemblance d'une trajectoire dans le modèle de la section 4

En temps discret, un individu (dont j'omets l'indice) effectue T transitions sur l'intervalle [0,T]. Les paramètres $K_1, K_2, p_1, p_2, p_3, r$ sont ceux indiqués dans le texte. La trajectoire de l'individu est résumée par les statistiques suivantes :

nmouv : nombre total de migrations sur la période (dont la totalisation sur tous les individus donne $M(T)$)

ns : nombre de migrations faisant sortir de l'état initial (dont la totalisation sur tous les individus donne $SM(T)$)

ne : nombre de migrations faisant rentrer dans l'état initial (dont la totalisation sur tous les individus donne $EM(T)$)

On en déduit le nombre de migrations entre deux états qui ne sont ni l'un ni l'autre l'état initial : $no = nmouv - (ne + ns)$ (dont la totalisation sur tous les individus donne $OM(T)$)

dini : temps d'attente de la première transition correspondant à une migration (si elle existe, c'est-à-dire si *nmouv>0*)

ditern : temps d'attente entre la première et la dernière migration à partir d'un état qui n'est pas l'état initial

diteri : temps d'attente entre la première et la dernière migration à partir d'un état qui est l'état initial

On en déduit le temps d'attente après la dernière mobilité :

$dap = T - (dini + ditern + diteri)$ dès lors que *nmouv>0.*

Si le nombre de migrations est nul, il peu s'agir ou bien d'un sédentaire, ou bien d'un mobile qui n'a pas encore bougé au bout de T transitions. La vraisemblance s'écrit : $K_1 + (1-K_1)(1-p_1)^T$

Si le nombre de migrations n'est pas nul, et vaut *nmouv>0*, il peut s'agir :

- soit d'un individu « nmouv-mobile » qui a été jusqu'au bout de ses mobilités, et ceci se produit avec une probabilité $(1-K_1)(1-K_2)^{nmouv-1}K_2$
- soit d'un individu « plus que nmouv-mobile » qui n'a pas encore fini ses migrations au bout de T transitions, et ceci se produit avec une probabilité $(1-K_1)(1-K_2)^{nmouv}$

Dans le premier de ces deux cas, la vraisemblance de la trajectoire est le produit de cette probabilité de tirage de l'individu par la quantité :

$(1-p_1)^{dini-1} p_1 r^{ne} {p_3}^{no} {p_2}^{ns-1} (1-p_2)^{diteri-(ns-1)} (1-(r+p_3))^{ditern-(ne+no)}$

Dans le second la vraisemblance de la trajectoire est le produit de la probabilité de tirage de l'individu par la même quantité, et encore par :

$(1-p_2)^{dap}$ si ne=ns

$(1-(r+p_3))^{dap}$ si ne=ns-1

Tous ces éléments permettent de calculer la vraisemblance d'une trajectoire quelconque.