Données longitudinales et modèles de survie

2. Concepts de base

André Berchtold

Département des sciences économiques, Université de Genève

Cours de Master

UNIVERSITÉ
DE GENÈVE

FACULTÉ DES SCIENCES
ÉCONOMIQUES ET SOCIALES
Département des sciences
économiques

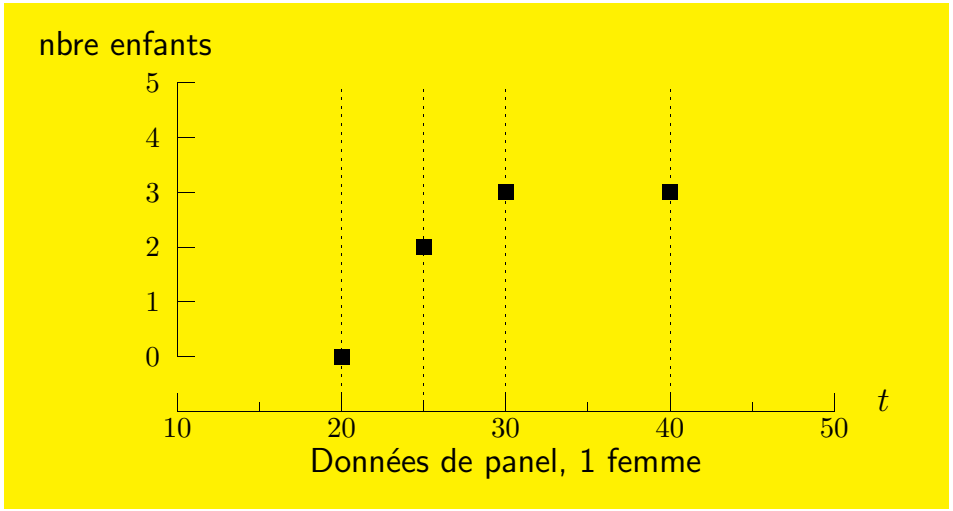
Plan du cours

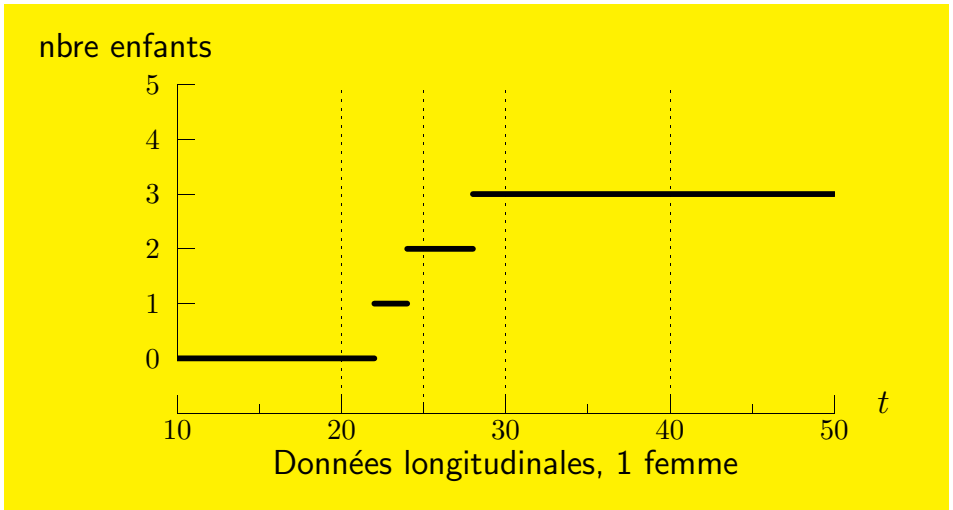
Plan du cours

Définition

- Data in which many units are observed over multiple time periods. (OECD Glossary of Statistical Terms)
- Synonyme courant, mais imprécis : Données de panel.
- Dans l'idéal, des données longitudinales sont des données pour lesquelles la valeur d'une (ou de plusieurs) variable d'intérêt est connue à chaque instant durant une période d'observation.

Panel versus longitudinal





Sources de données longitudinales

- **Suivis individuels :** Chaque événement important est enregistré lorsqu'il survient.
  - Dossier médical
- **Panels :** Observation périodique des mêmes individus.
  - Panel Suisse des Ménages (PSM)
  - Transitions de l'Ecole à l'Emploi (TREE)
- **Enquête rétrospective :** Fait appel à la mémoire des interviewés.
  - Enquête biographique rétrospective du PSM
- **Couplage des données de diverses sources :** Recensements successifs, données fiscales, registre des population, actes de mariage, actes de décès, ...
  - Wanner & Delaporte (2001) : recensements et état-civil
  - Perroux & Oris (2005) : Genève au XIXème (1816 - 1843), recensements, actes de mariage et de décès, immigration

Types de données longitudinales

- **Séquences :** Suites chronologiques d'états ou d'événements.
  - Séquences d'états
    - ..., célibataire à 25, marié à 26, marié à 27, divorcé à 28, ...

    Les panels génèrent naturellement des séquences d'états.
  - Séquences d'événements
    - début d'une union, 1er enfant, mariage, 2ème enfant, chômage, divorce, ...
- **Evénements datés :** Suites d'événements avec indication de leur date.
  - fin de l'école secondaire à 19 ans, premier emploi à 19 ans, première union à 20 ans, 1er enfant à 23 ans, ...

On peut en général passer d'une présentation à l'autre

$\rightarrow$ question d'organisation des données

Représentation des données (1)

Evénements datés

fin école secondaire en 1970	premier job en 1971	mariage en 1973

Séquences d'états

année	1969	1970	1971	1972	1973
état civil	célib.	célib.	célib.	célib.	marié
niv. formation	primaire	secondaire	secondaire	secondaire	secondaire
emploi	non	non	1er	1er	1er

Episodes

id	de	à	état civil	formation	emploi
1	1969	1969	célib.	primaire	non
1	1970	1970	célib.	secondaire	non
1	1971	1972	célib.	secondaire	1er
1	1973	1973	marié	secondaire	1er

Représentation des données (2)

Données personnes-périodes

id	année	état civil	formation	emploi
1	1969	célib.	primaire	non
1	1970	célib.	secondaire	non
1	1971	célib.	secondaire	1er
1	1972	célib.	secondaire	1er
1	1973	marié	secondaire	1er
2	1969	célib.	primaire	1er
2	1970	célib.	primaire	1er
2	1971	marié	primaire	1er
2	1972	marié	primaire	non
2	1973	marié	primaire	non
⋮				

Durées

- L'analyse des biographies s'intéresse essentiellement à la durée des épisodes ou alternativement au risque que l'épisode se termine après une durée t.
- La **durée de séjour** dans un état est le temps s'écoulant entre deux événements (durée des épisodes).
- Exemples : temps entre
  - immatriculation et obtention du diplôme
  - premier engagement et premier changement d'emploi
  - mariage et 1ère naissance
  - mariage et divorce
  - diagnostic d'une maladie grave et décès

Calendrier et cohortes

- Les **dates** sont des mesures de type intervalle (pas de zéro absolu) dont l'**origine est arbitraire**. On peut mesurer le temps depuis une date fixe (par exemple depuis le 1er janvier 1900) ou relative (date de naissance, début de l'emploi).
  $\rightarrow$ plusieurs calendriers (horloges) possibles pour un même individu
- Exemples :
  - âge : durée depuis la naissance
  - durée mariage : durée depuis le mariage
  - époque de l'événement : durée depuis une date fixe telle que le début du siècle
- **Cohorte** : Ensemble d'individus ayant vécu un événement particulier (naissance par exemple) en même temps.

Fichiers de données (1)

- **Fichier individus** : Une ligne par individu (durée en mois).

	emploi					mariage					...
indiv	déb 1	fin 1	déb 2	fin 2	...	déb 1	fin 1	déb 2	fin 2	...	...
1	204	216	260	350	...	300	-	-	-	...	...
2	240	400	401	-	...	340	500	-	-	...	...
⋮											

Fichiers de données (2)

- **Fichier épisodes** : Une ligne par épisode (nouvelle ligne chaque fois qu'un événement se produit).

indiv	début	fin	emploi	nbre emplois	marié	···
1	1	203	non	0	non	···
1	204	216	oui	1	non	···
1	217	259	non	1	non	···
1	260	299	oui	2	non	···
1	300	350	oui	2	oui	···
2	1	239	non	0	non	···
2	240	339	oui	1	non	···
2	340	400	oui	1	oui	···
2	401	500	oui	2	oui	···
2	···					
⋮						

Fichiers de données (3)

- **Fichier personnes-périodes** : Pour chaque personne, une ligne pour chaque période où elle est observée.

indiv	mois	emploi	nbre emplois	marié	···
1	1	non	0	non	···
⋮	⋮	⋮	⋮	⋮	
1	203	non	0	non	···
1	204	oui	1	non	···
⋮	⋮	⋮	⋮	⋮	
1	216	oui	1	non	···
1	217	non	1	non	···
⋮	⋮	⋮	⋮	⋮	
1	259	non	1	non	···

Fichiers de données (4)

indiv	mois	emploi	nbre emplois	marié	···
1	260	oui	2	non	···
⋮	⋮	⋮	⋮	⋮	
1	299	oui	2	non	···
1	300	oui	2	oui	···
⋮	⋮	⋮	⋮	⋮	
1	350	oui	2	oui	···
2	1	non	0	non	···
⋮	⋮	⋮	⋮	⋮	
2	239	non	0	non	···
2	240	oui	1	non	···
⋮	⋮	⋮	⋮	⋮	

Représentation des épisodes (1)

- Méthode la plus générale : (o_i, d_i, s_i, r_i), où
  - o_i = état d'origine (définissant le début de l'épisode)
  - d_i = état de destination (un changement vers cet état définit la fin de l'épisode)
  - s_i = date de début de l'épisode (date à laquelle l'état d'origine est observé pour la première fois)
  - r_i = date de fin de l'épisode (date à laquelle l'état de destination est observé pour la première fois)
- Exemple pour l'emploi de l'individu 1.

épisode i	o_i	d_i	s_i	r_i
1	sans emploi	employé	1	204
2	employé	sans emploi	204	217
3	sans emploi	employé	217	260
⋮				

Représentation des épisodes (2)

- Représentation simplifiée : (t_i, e_i) où
  - $t_i = r_i - s_i$ durée de l'épisode
  - $e_i = 1$ si l'événement a lieu en fin d'épisode et 0 sinon
- Ne considère que des événements simples (oui/non) et la durée de l'épisode.
- Exemple pour l'événement "fin de l'emploi".

épisode i	t_i	e_i
1	203	0
2	13	1
3	43	0
⋮		

Représentation des épisodes (3)

- Le type de représentation dépend du logiciel utilisé.
- SPSS par exemple n'autorise que la représentation simplifiée (t_i, e_i).
- Le package *survival* de R donne plus de souplesse en permettant d'utiliser plusieurs représentations, dont notamment la représentation simplifiée (t_i, e_i) et une représentation $(t1_i, t2_i, e_i)$, où
  - $t1_i$ = date de début de l'épisode
  - $t2_i$ = date de fin de l'épisode (**avant** la survenue du changement d'état)
  - $e_i = 1$ si l'événement a lieu en fin d'épisode et 0 sinon

Représentation des épisodes (4)

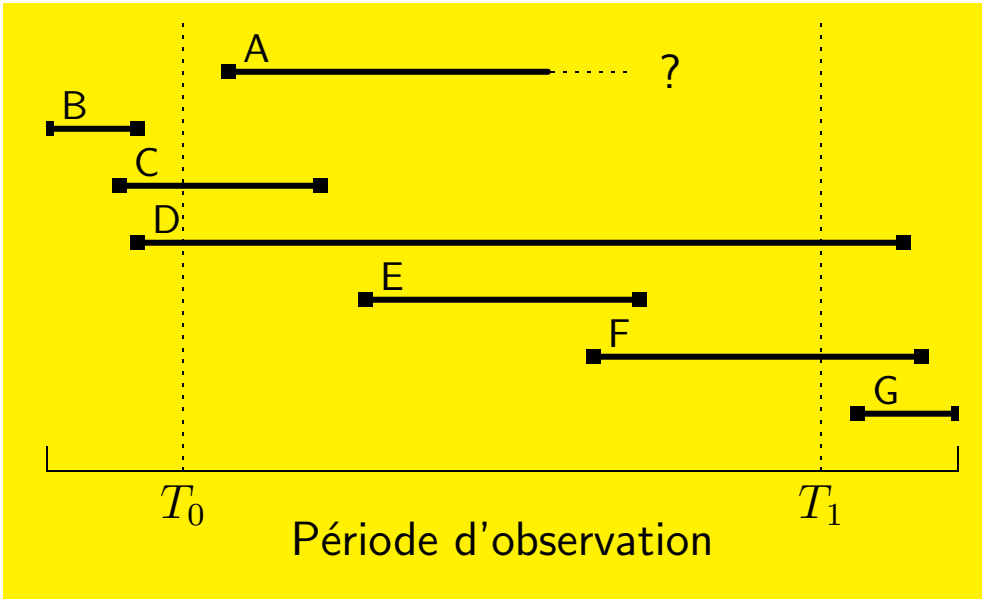
- Exemple pour l'événement "fin de l'emploi".

épisode i	$t1_i$	$t2_i$	e_i
1	1	203	0
2	204	216	1
3	217	259	0
⋮			

- Par extension, il est possible d'utiliser ce type de représentation avec une variable supplémentaire indiquant le type de censure des données, ce qui permet notamment d'utiliser des données censurées à gauche.

Définition

- Données pour lesquelles l'événement (initial ou terminal) n'a pas lieu durant la période observée.
- Exemple d'un étudiant
  - qui a interrompu ses études (A)
  - dont on ne connaît pas la date de début d'études (C, D)
  - toujours en études (D, F)



Traitement

- Les données censurées requièrent un traitement spécial.
- On ne peut pas simplement ignorer ces cas (que l'on sait être exposés au risque durant la période observée).
  - Données tronquées à droite : dont on ne connaît pas la date de fin de l'épisode.
  - Données tronquées à gauche : dont on ne connaît pas la date de début de l'épisode.

  $\rightarrow$ Durée de séjour effective plus longue que la durée observée.
- Exclure ces cas conduit à sous-estimer la durée moyenne de séjour.
- On sait bien traiter les données tronquées à droite, mais le traitement des données tronquées à gauche est plus délicat (impossible avec la représentation (t_i, e_i)).

Exemple de biais

Age au mariage en Suisse, femmes

Année de naissance	moyenne	écart type	médiane	maximum	*n*
1935 et avant	26.0	5.11	25	53	302
1936 - 1955	24.7	4.61	24	50	852
1956 - 1970	26.0	4.32	26	43	773
1971 et plus tard	24.3	2.72	24	30	126
Total	25.4	4.54	25	53	2053

Source : PSM, enquête biographique de 2001/02

- Age maximum des femmes de la dernière classe = 30 ans.
  $\rightarrow$ que des femmes mariées jeunes
  $\rightarrow$ moyenne, médiane et dispersion sous-estimées

Plan du cours

Trois grandes catégories d'événements

- **Evénement unique** : Dans le domaine biomédical ou industriel, l'événement est le plus souvent unique (la mort, la rupture d'une pièce, ...)
  $\rightarrow$ analyse de survie
- **Evénements multiples** : En sciences sociales, en démographie, on analyse souvent conjointement plusieurs événements (fin de scolarité, premier emploi, second emploi, mariage, naissance du premier enfant, du deuxième, divorce, veuvage, mort).
  $\rightarrow$ étude des liens entre les risques (risques compétitifs, interactions entre risques, Courgeau & Lelièvre, 1989)
- **Evénement à répétition** : Parfois, un même événement peut se produire à plusieurs reprises (mariage, naissance d'un enfant, changement d'emploi).

Evénements multiples (1)

- Les événements multiples peuvent être classifiés en différents types.
- **I.** Evénements dont le type est déterminé par un processus causal distinct de l'occurence de l'événement.
  On décide de partir en vacances (événement), puis on décide de l'endroit des vacances (type de l'événement).

Evénements multiples (2)

- **II.** Evénements causés par des processus distincts.
  - **IIa.** La survenance d'un événement implique la non-survenance d'un autre (risques compétitifs).
    Si l'on décède d'un cancer, on ne décédera pas d'un accident de la circulation.
  - **IIb.** Evénements rendant impossible l'observation d'un autre événement.
    Un employé prenant sa retraite ne peut plus être licencié.
  - **IIc.** Evénements totalement indépendants les uns des autres.
    Avoir un accident de la circulation et aimer la littérature.
  - **IId.** Evénements dont la survenance augmente ou diminue (mais pas totalement) les chances d'occurence d'un autre événement.
    Terminer ses études augmente la probabilité de commencer un emploi.

Plan du cours

Décrire la durée d'un épisode

- La durée T d'un épisode (qui correspond aussi au temps avant la survenance de l'événement dans le cas d'événements uniques) est une variable aléatoire.
- Elle est décrite par plusieurs fonctions liées les unes aux autres.
- Ces fonctions peuvent être définies indifféremment en temps continu ou en temps discret.

Fonction de densité

- Cette fonction représente la probabilité que l'événement survienne à chaque temps t (cas discret) ou dans chaque intervalle de temps (cas continu).
- En temps continu, la fonction de densité $f(t)$ de la durée T s'écrit

$$f(t) = \lim_{dt \to 0} \frac{P(T \in [t, t + dt])}{dt}$$

- En temps discret, nous avons une distribution de probabilité définie par

$$p_t = P(T = t)$$

Fonction de répartition ou de distribution

- Cette fonction représente la probabilité que l'événement se soit déjà produit au temps t.
- En temps continu, la fonction de répartition $F(t)$ de la durée T s'écrit

$$F(t) = P(T \leq t) = \int_0^t f(t)\, dt$$

- En temps discret, nous avons une distribution de probabilité cumulée définie par

$$F_t = \sum_{i=1}^{t} p_i$$

Fonction de survie ou de séjour

- Cette fonction représente la probabilité que l'événement ne se soit pas encore produit au temps t.
- En temps continu, la fonction de survie $S(t)$ s'écrit

$$S(t) = 1 - F(t)$$

- En temps discret, nous avons

$$S_t = 1 - F_{t-1}$$

Fonction de hasard ou de risque instantané

- Cette fonction décrit la densité (proportion) de cas subissant l'événement en t parmi ceux exposés au risque en t.
- En temps continu, la fonction de hasard $h(t)$ s'écrit

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T \leq t + dt | T \geq t)}{dt} = \frac{f(t)}{S(t)}$$

- En temps discret, nous avons

$$h_t = \frac{p_t}{S_t}$$

Fonction de hasard cumulé

- Cette fonction décrit le nombre moyen d'événements qui serait observé si l'individu étudié était exposé en permanence au risque.
- En temps continu, la fonction de hasard cumulé $H(t)$ s'écrit

$$H(t) = \int_0^t h(t)\, dt$$

- En temps discret, nous avons

$$H_t = \sum_{i=1}^{t} h_i$$

Exemple (1)

- Personnes quittant leur premier emploi. Suivi de 200 personnes durant une période de 5 années (temps discret).

t	# restant	# quittant	p_t	F_t	S_t	h_t
1	200	11	0.055	0.055	1	0.055
2	189	25	0.125	0.180	0.945	0.132
3	164	10	0.050	0.230	0.820	0.061
4	154	13	0.065	0.295	0.770	0.084
5	141	12	0.060	0.355	0.705	0.085
>5		129	0.645	1	0.645	1
total	848	200				

Exemple (2)

- La probabilité de quitter son emploi après 2 ans se calcule comme

$$p_2 = \frac{\#\text{ quittant en } t = 2}{\#\text{ total}} = \frac{25}{200} = 0.125$$

- La probabilité de quitter son emploi au plus tard après 2 ans se calcule comme

$$F_2 = \frac{\#\text{ quittant en } t \leq 2}{\#\text{ total}} = \frac{11 + 25}{200} = 0.180$$

Exemple (3)

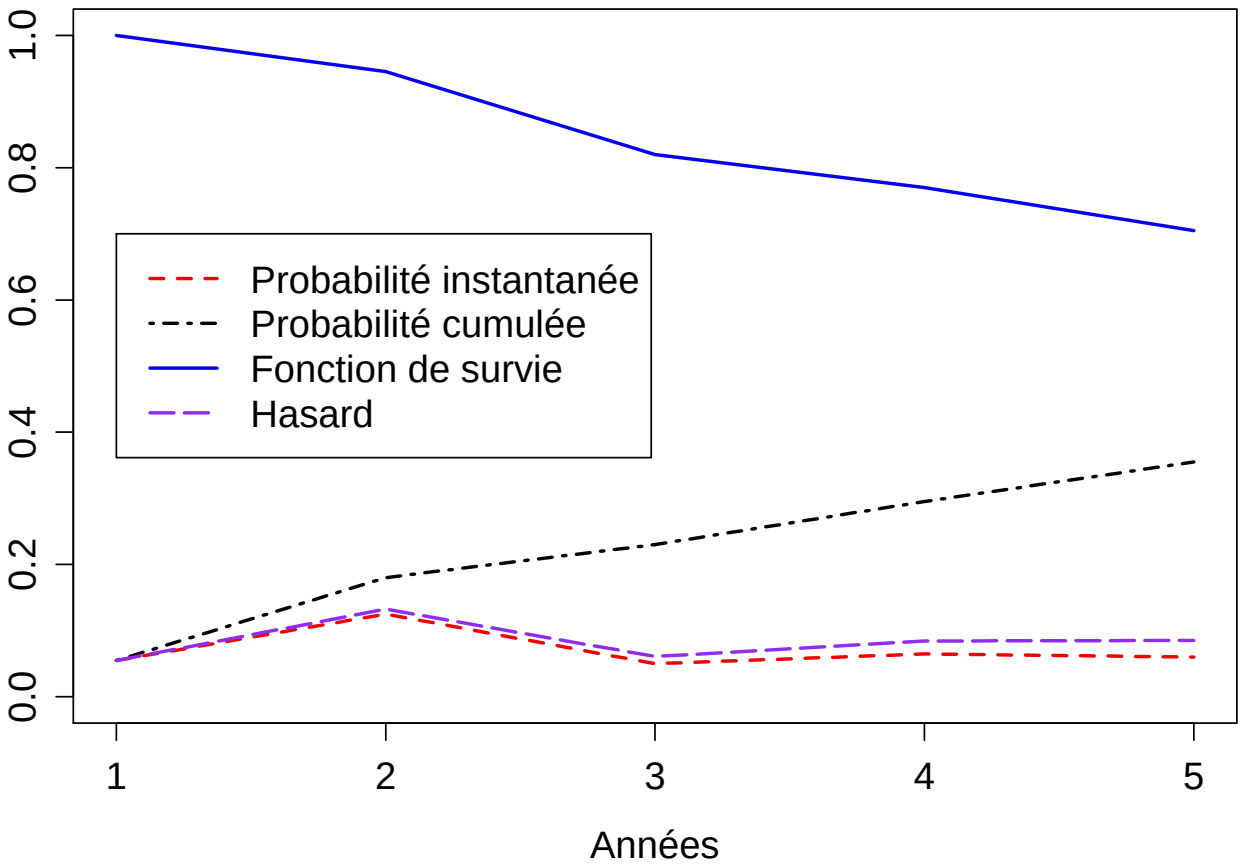
- La probabilité de ne pas avoir quitté son emploi au début de la deuxième année se calcule comme

$$S_2 = 1 - F_1 = 1 - 0.055 = 0.945$$

- Le risque (instantané) de quitter son premier emploi après 2 ans sachant qu'on a gardé son emploi la première année se calcule comme

$$h_2 = \frac{\#\text{ quittant en } t = 2}{\#\text{ restant en } t = 2} = \frac{25}{189} = 0.132 \quad = \quad \frac{p_2}{S_2} = \frac{0.125}{0.945}$$

Exemple (4)


1.0
0.8
0.6
0.4
0.2
0.0
Probabilité instantanée
Probabilité cumulée
Fonction de survie
Hasard
1
2
3
4
5
Années

Exemple (5)

- L'exemple précédent est un peu simpliste dans la mesure où l'on suppose implicitement que toutes les personnes quittant leur emploi une année donnée le quitte à la fin de celle-ci.
- La méthode actuarielle permet de corriger ceci en répartissant les événements tout au long de l'année.

Bibliographie

- Courgeau D, Lelièvre É (1989) *Analyse démographique des biographies.* Paris : Editions de l'INED.
- Perroux O, Oris M (2005) Présentation de la base de données de la population de Genève de 1816 à 1843. Séminaire statistique sciences sociales, Université de Genève.
- Wanner P, Delaporte E (2001) Reconstitution de trajectoires de vie à partir des données de l'état civil (BEVNAT). Une étude de faisabilité. Rapport de recherche, Forum Suisse des Migrations.