Les trajectoires d'investissement d'exploitations laitières en Ille-et-Vilaine

François Legland

Rapport du stage d'application en statistique de 2e année

François LEGLAND

Structure d'accueil :

AGROCAMPUS OUEST
UMR SMART-LERECO

Thème du stage :

Analyse des trajectoires d'investissement d'exploitations laitières en Ille-et-Vilaine

Lieu
Agrocampus Ouest
65 rue de Saint-Brieuc
35042 Rennes

Maîtres de stage
Aude Ridier
Laure Latruffe

Tuteur pédagogique
Ronan Le saout

Promotion 2019

7 septembre 2018

Remerciements

Je tiens à remercier tout particulièrement Aude Ridier et Laure Latruffe pour la confiance qu'elles m'ont accordée durant ce stage et pour les corrections qu'elles ont apportées à mon rapport. Mon appropriation au sujet n'aurait pas été aussi rapide sans leur précieuse expertise et pédagogie.

Un grand merci à Loïc Lévi pour sa disponibilité, son écoute, et ses réponses lors de mes différents questionnements au cours de ce stage.

J'adresse finalement mes sincères remerciements à tout le personnel de l'unité SMART-LERECO pour son accueil chaleureux.

Table des matières

Cadre du stage

Mes deux tutrices de stage ont été Aude Ridier et Laure Latruffe, respectivement enseignant-chercheur à Agrocampus Ouest et chercheur à l'INRA. Agrocampus Ouest est une école d'ingénieurs publique spécialisée dans l'enseignement supérieur et la recherche en sciences agronomiques, agroalimentaires, horticoles et du paysage. L'Institut National de Recherche Agronomique (INRA) est un établissement de recherche public, dépendant à la fois du ministère chargé de l'Agriculture et du ministère chargé de la Recherche.

Mon stage s'est déroulé au sein de l'UMR SMART-LERECO, localisée à Rennes et dont font toutes deux partie mes tutrices. Les études menées au sein de cette unité de recherche en économie portent sur le comportement des exploitants agricoles, leurs prises de décisions dans un cadre statique ou dynamique et l'impact des politiques publiques sur ce comportement.

Aude Ridier est titulaire d'une thèse en économie rurale. Maître de conférence en finance d'entreprise, ses travaux recouvrent différents aspects de l'activité agricole dont : le financement, la gestion, les choix d'investissements et les pratiques de production.
Laure Latruffe est directrice de recherche en microéconomie de la production agricole. Ses sujets de recherche ont principalement trait à la performance, productivité et viabilité des exploitations, la transmission des exploitations ou encore l'impact de politiques publiques.

Ma mission s'inscrivait dans le cadre d'un partenariat entre Agrocampus Ouest et le Crédit Agricole en Bretagne, concrétisée par la Chaire "Entreprises et Economie Agricole" (EEA). D'un point de vue de la recherche, celle-ci a notamment pour objectif de "partager les approches pragmatiques du Crédit Agricole et les outils d'analyse développés par les chercheurs pour mieux répondre aux questions et enjeux de demain" [1]. Cette Chaire finance une thèse portant sur les comportements et performances des exploitations agricoles selon la position dans leur cycle de vie, réalisée par Loïc Lévi. Mon travail consistait en l'exploitation statistique des données utilisées dans cette thèse.

1. Voir http ://chaire-eea.agrocampus-ouest.fr

Introduction

En économie, la théorie identifie des périodes distinctes dans la vie d'une entreprise. Ces périodes correspondent successivement à un mouvement d'installation, d'expansion, de survie puis de disparition ou de transmission. Les entreprises du secteur agricole, c'est-à-dire les agriculteurs, passent également par ces phases du cycle de vie. Les exploitations agricoles peuvent être de deux formes juridiques : des exploitations individuelles, c'est-à-dire avec un seul chef d'exploitation ; des exploitations du type sociétaire (par exemple GAEC) où le management est assuré par plusieurs associés. Dans le premier cas, le cycle de vie de l'exploitation suit généralement le cycle de vie personnel du chef d'exploitation. Dans le deuxième cas, l'arrivée d'un nouvel associé correspond à une phase d'installation. Une des particularités des exploitations agricoles, notamment en élevage, est le fort besoin en capital : foncier (c'est-à-dire les terres), bâtiments, cheptel. L'investissement au cours du cycle de vie, et notamment lors de la phase d'installation, est donc important.

Les quelques années suivant l'installation sont donc cruciales pour une exploitation : besoin en capital important mais aussi nécessité de se familiariser avec la technologie, etc. Ses résultats (en termes de revenus, de performances) peuvent donc s'en trouver affectés. La littérature identifie le seuil des cinq ans après installation comme particulièrement pertinent pour délimiter cette période.
La première mission de ce stage était de mettre en évidence la particularité de ces cinq premières années sur la base de différents indicateurs.
Au-delà de cette comparaison entre nouveaux installés et exploitations plus anciennes, c'est plus largement la notion de trajectoire individuelle (c'est-à-dire l'évolution au sein de chaque exploitation) qu'il s'agissait de traiter.

Les données utilisées sont des données comptables annuelles. Elles forment un panel non cylindré d'environ 4 000 exploitations laitières d'Ille-et-Vilaine, observées sur la période 2005-2014. Les données comptables fournissent sur les exploitations à la fois un relevé annuel de leur patrimoine mais aussi des charges et produits relatifs à leur activité. Des informations concernant la date de création de l'exploitation et l'âge de ou des exploitant(s) nous ont également été remises.

Après un travail préalable de mise en forme et de "nettoyage" des données, une première partie du stage a été consacrée à la comparaison entre exploitations dans leur phase d'installation/exploitations au-delà de cette phase, avec parfois une distinction supplémentaire suivant le statut juridique de celles-ci. La première partie du rapport retrace ces travaux.
Dans un second temps nous nous sommes plus particulièrement intéressés aux trajectoires des exploitations, c'est-à-dire à leur évolution dans le temps cette fois-ci sans découpage en plusieurs phases. Notre objectif était surtout d'appréhender l'hétérogénéité des trajectoires. La seconde partie du rapport a pour but d'introduire et de mettre en applications les méthodes statistiques adaptées à une telle analyse.

Chapitre 1

Effets de l'installation et du statut juridique dans un cadre statique

Démarche

Pour éviter le "problème" de dépendance de nos observations dans nos données de panel, on va dans un premier temps raisonner dans un cadre statique et mettre en partie de côté l'aspect dynamique du phénomène observé. Comment ? Pour chaque variable d'intérêt et pour chaque exploitation, on ne va non pas s'intéresser directement à l'ensemble des valeurs prises chaque année mais plutôt à une unique valeur agrégée, calculée sur cet ensemble (ex : la moyenne intertemporelle, le taux de croissance observé, etc.). Malgré ses apparentes limites, cette méthode dite de *Derived variables analysis* [1], comporte deux avantages non négligeables. Le premier est qu'elle permet de se ramener à des méthodes d'analyse statistique classiques, où l'hypothèse d'indépendance de nos observations est vérifiée : à une exploitation correspondra en effet une unique observation, et l'indépendance inter-exploitations restera elle toujours supposée. Le second avantage est lié au premier puisqu'il réside dans la facilité d'interprétation des résultats alors obtenus.
La première partie de ce chapitre propose une analyse exploratoire des données "statiques", où l'on s'intéresse notamment à l'effet des facteurs "nouvel installé" et "statut juridique" sur différent résultats. Dans un but d'affiner notre inférence sur le facteur "nouvel installé", nous proposons finalement une méthode d'estimation par matching. L'idée est de comparer chaque exploitation nouvellement installée avec une exploitation plus ancienne, mais qui lui est relativement semblable par rapport à des critères prédéfinis. En procédant ainsi, notre intérêt est notamment de pouvoir contrôler d'éventuels biais d'estimations liés par exemple à la conjoncture ou à la structure des exploitations.

1.1 Premières comparaisons

Quelle *derived variable* ?

Comme expliqué plus haut, nous allons dans ce chapitre nous ramener à l'analyse de *derived variables*, c'est-à-dire à des variables censées résumer au mieux l'information enregistrée sur toute la période. Le choix d'une variable plutôt qu'une autre relève de ce que l'on cherche à étudier.

Par exemple, comme notre objectif premier est de capter d'éventuelles différences de comportement en matière d'investissements, il a dans ce cas été jugé pertinent de mesurer celles-ci sur la base du *taux de croissance annuel composé* (CAGR [2] : *Compound Annual Growth Rate* en anglais) du capital de l'exploitation, représenté par son niveau d'immobilisations nettes hors foncier (*immo* dans le dictionnaire des variables [3]). Cet indicateur présente différents avantages :

– les exploitations ne sont pas toutes propriétaires du terrain qu'elles exploitent ; le cas échéant on parle de *fermage*. Si elle utilise le fermage, une exploitation affiche un montant comptable de

1. Voir Chapitre 6 dans Diggle et al.(2002)[2]
2. À distinguer du *taux de croissance annuel moyen* (AAGR : *Average Annual Growth Rate*)
3. Toutes les variables citées dans le rapport sont définies dans le tableau A.1 en annexe A

biens foncier quasiment nul. Celui-ci n'est alors pas représentatif du capital réellement utilisé. C'est également le cas si l'exploitation est propriétaire de terrains qu'elle n'utilise pas pour son activité. Il a donc été décidé ici de ne pas inclure les immobilisations foncières dans *immo*, variable qui désigne ce que nous appellerons à chaque fois "capital" ; *immo* correspond donc à la somme : du montant des immobilisations nettes de matériel ($immo\_mat\_net$), du montant des immobilisations nettes de construction ($immo\_const\_net$) et du montant net des biens vivants ($biens\_vivants\_net$).

- il quantifie bien l'intensité avec laquelle une exploitation a investi/désinvesti durant sa période d'observation. Le CAGR, calculé entre deux années, s'interprète comme le taux de croissance annuel constant auquel aurait crû/décru le capital pour obtenir la différence observée entre ces deux instants. Formellement, il se définit ici comme τ_i tel que $Y_{i,k_i} = (1+\tau_i)^{k_i}.Y_{i,1}$ où i désigne l'exploitation, $Y_{i,1}$ le niveau initial de capital (lorsque i est observée pour la première fois), Y_{i,k_i} le niveau final de capital (lorsque i est observée pour la dernière fois) (avec $k_i \in \{2, ..., 10\}$).
- contrairement au taux de croissance calculé entre la première et dernière année d'observation, il est insensible à la durée d'observation de l'exploitation. Prenons l'exemple de deux exploitations, à niveaux égaux de capital initial et d'investissement annuels. Si l'une est observée plus longtemps, son taux de croissance affichée est naturellement le plus élevée ; cette différence est corrigée avec le CAGR.
- à la différence de l'AAGR, il peut être calculé malgré un manque d'observation une année intermédiaire. De la définition donnée plus haut il peut en effet être aisément déduit que $\tau_i = \left(\frac{Y_{i,k_i}}{Y_{i,1}}\right)^{\frac{1}{k_i}} - 1$.

1.1.1 ANOVA à deux facteurs sur l'échantillon complet

Dès que l'on souhaite mesurer l'effet d'un ou plusieurs facteurs sur une variable continue, une approche communément empruntée est celle de l'Analyse de la Variance (ANOVA). Nous ne ferons donc pas figure d'exception en étudiant selon cette méthode l'effet des facteurs "nouvel installé" et "statut juridique" sur le CAGR du niveau d'immobilisations nettes (hors foncier) (*immo*).
L'ANOVA peut être vue comme une généralisation du test de Student (t-test) de comparaison de moyennes à deux groupes. Les groupes implicites sont donnés par croisement des facteurs retenus. Dans notre cas, ces groupes sont les suivants :

— *IndivCreaInf* : désigne les exploitations individuelles observées dans leurs cinq premières années après installation (<=5 ans)

— *IndivCreaSup* : désigne les exploitations individuelles observées au-delà de cinq ans après installation

— *SocCreaInf* : désigne les exploitations en société observées dans leurs cinq premières années après installation (<= 5 ans)

— *SocCreaSup* : désigne les exploitations en société observées au-delà de cinq ans après installation.

Notons que les exploitations observées avant *et* après leur cinquième année de création sont classées en tant que *creaInf* et que l'information fournie après leur cinquième année d'existence n'est pas prise en compte.

Préliminaires

Le tableau B.1 en annexe B.0.1 rassemble des indicateurs calculés sur l'ensemble de l'échantillon. On y observe que le CAGR individuel moyen s'élève à -1.47%, avec un écart-type de 14.14 points. Alors qu'un quart des exploitations présentent un CAGR individuel supérieur à 4.16%, un autre quart de celles-ci révèlent un CAGR individuel inférieur à -7.89%.

Une étape préliminaire à l'ANOVA consiste à représenter la répartition des CAGR au sein de chaque ensemble. La figure 1.1 ne montre pas de différence très marquée entre ces répartitions. Elles semblent toutes à peu près centrées sur la valeur $\tau_0 = 0$, c'est-à-dire le cas où le niveau d'immobilisations individuel reste constant dans le temps. On note la présence de nombreuses valeurs atypiques (points

au-delà des valeurs adjacentes), caractéristiques de comportements "extrêmes" en termes d'investissement/désinvestissement.

Écriture du modèle

Notre modèle ANOVA à deux facteurs avec effet d'interaction s'écrit de la façon suivante :

$$\tau_i = \mu + \alpha \mathbb{1}_{SOC_i} + \beta \mathbb{1}_{crea\_Sup_i} + \gamma \mathbb{1}_{SOC_i}.\mathbb{1}_{crea\_Sup_i} + \epsilon_i$$

où : les ϵ_i sont supposés i.i.d $\mathcal{N}(0, \sigma^2)$; $\mathbb{1}_{SOC_i}$ vaut 1 si *statut* vaut "Soc", 0 sinon ; $\mathbb{1}_{crea\_Sup_i}$ vaut 1 si *pos_crea* vaut "creaSup", 0 sinon. [4]

FIGURE 1.1 – Intragroup repartition of *immo*'s CAGR

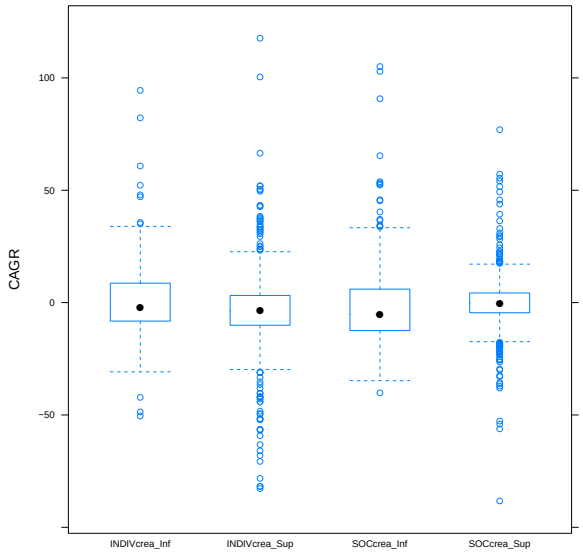


TABLE 1.1 – Two factor ANOVA with interaction effect on *immo*'s CAGR

	Dependent variable :
	τ
$\hat{\alpha}$	-2.304^{*}
	(1.288)
$\hat{\beta}$	-5.577^{***}
	(1.079)
$\hat{\gamma}$	5.736^{***}
	(1.412)
$\hat{\mu}$	1.852^{*}
	(0.992)
Observations	2,847
R^2	0.017
Adjusted R^2	0.016
Residual Std. Error	14.030 (df = 2843)
F Statistic	16.609^{***} (df = 3 ; 2843)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

4. Voir section A.0.3 pour la définition des variables en question et leurs modalités

TABLE 1.2 – Mean *immo*'s CAGR by group

	IndivCreaInf	*IndivCreaSup*	*SocCreaInf*	*SocCreaSup*
Number of farms	200	1092	292	1264
Estimated mean level	$\hat{\mu} = 1.852$	$\hat{\mu} + \hat{\beta} = -3.725$	$\hat{\mu} + \hat{\alpha} = -0.452$	$\hat{\mu} + \hat{\alpha} + \hat{\beta} + \hat{\gamma} = -0.293$

Résultats

Le tableau 1.1 présente les résultats du modèle estimé et le tableau 1.2 les niveaux moyens estimés par groupe qui en découlent. On y observe que le CAGR moyen des exploitations individuelles de moins de 5 ans, $\hat{\mu} = 1.185$, est significativement différent de 0. L'effet estimé du statut sociétaire pour les exploitations nouvellement installées, $\hat{\alpha} = -2.304$, est significativement négatif. L'effet estimé du statut sociétaire pour les exploitations déjà installées $\hat{\alpha} + \hat{\gamma} = 3.432$ est lui positif et significativement différent de $\hat{\alpha}$. Pour les exploitations individuelles, le fait d'être déjà installé diminue en moyenne de 5.57 points le CAGR observé sur la période. Parmi les exploitations en société, l'effet est inverse puisqu'en moyenne les exploitations déjà installées ont un CGAR supérieur de $\hat{\beta} + \hat{\gamma} = 0.159$ points aux exploitations nouvellement installées.
Le très faible pouvoir explicatif du modèle ainsi que les violations constatées aux hypothèses d'homoscedasticité et de normalité des termes d'erreur empêchent cependant toute interprétation causale de ces résultats. Cela confirme avant tout que les déterminants à l'investissement sont multiples et complexes.

1.1.2 Comparaison avant/après cinq ans pour les nouveaux installés

Dans la section précédente, nous avons étudié l'impact du statut "nouvel installé" sur le CAGR observé de façon relativement simple. Il va s'agir maintenant d'exploiter a minima le caractère longitudinal de nos données, en raisonnant par différenciation inter-temporelle. Plus précisément la méthodologie qui va suivre prend ses racines dans une idée communément émise en données répétées : l'existence d'effets individuels inobservés.
Formellement ces effets correspondent à l'idée que pour toute observation j faite sur l'individu i, il existe un niveau c_i tel que l'on ait : $y_{ij} = \beta_0 + \beta_1 x_{ij} + c_i + \epsilon_{ij}$, avec y et x des variables quelconques et avec les hypothèses usuelles sur les termes d'erreur ϵ_{ij} (indépendants, centrés, homoscédastiques et normaux).
Comme c est inobservé, toute estimation par moindres carrés ordinaires de l'effet β_1 de x_{ij} suivant le modèle donné plus haut est soumise à un potentiel biais de variable omise ; c'est notamment le cas si $cov(x_{ij}, c_i) \neq 0$. L'alternative proposée, qui est d'ailleurs celle des estimateurs *first-difference* ou *within* en données de panel, est donc de raisonner par différenciation, en considérant par exemple plutôt l'écriture $y_{ij+1} - y_{ij} = \beta_1(x_{ij+1} - x_{ij}) + \epsilon_{ij+1} - \epsilon_{ij}$.
C'est cette logique que nous allons suivre, pour différents y et avec $x_{ij} = 1_{crea\_Sup_{ij}}$. Nous considérons ici uniquement les exploitations observées avant et après leur cinquième année d'existence (cela représente 6% de notre échantillon total) et raisonnons pour chacune selon deux périodes $j \in \{1, 2\}$ ($j = 1$ lorsque l'exploitation a moins de 5 ans, $j = 2$ lorsque l'exploitation a 6 ans ou plus).

Méthodologie

Parce que nous n'introduisons pas d'autre covariable que x_{ij} (unidimensionnel), notre analyse peut se suffire à un test de comparaison de moyennes sur échantillons appariés.
Notons $\forall i \in \{1, ..., N\}$:$D_i = (y_{i2} - y_{i1})$, les D_i sont supposés i.i.d d'espérance μ_D et de variance inconnue σ_D^2. Comme $\forall i : x_{i2} - x_{i1} = 1$, on a en fait que $\mu_D = \beta_1$. [5] L'hypothèse testée est celle d'un effet nul du passage des cinq ans sur le niveau espéré d'output y :

$$H_0 : \mu_D = 0 \quad vs \quad H_1 : \mu_D \neq 0$$

Pour un échantillon assez grand (ici, $N = 242$), la statistique de test $z - value = \sqrt{N}\frac{\overline{D}_N}{\hat{\sigma_D}}$ suit approximativement une loi normale centrée réduite.

5. Par linéarité de l'espérance et comme les termes d'erreur sont centrés : $\mathbb{E}(D_i) = \mathbb{E}(\beta_1(x_{ij+1} - x_{ij}) + \epsilon_{ij+1} - \epsilon_{ij}) = \beta_1$

Nous avons retenu deux types de *derived variable*[6] y : le niveau moyen et le CAGR. Par exemple, lorsque y est égale au niveau moyen pour *output_lait* on aura : y_{i1} = le niveau annuel moyen de *output_lait* lorsque l'exploitation i a moins de 5 ans, y_{i2} = le niveau annuel moyen de *output_lait* lorsque l'exploitation a 6 ans ou plus. Lorsque y est égale au CAGR de *output_lait* on aura : y_{i1} = CAGR d'*ouput_lait* calculé lorsque l'exploitation i a moins de 5 ans, y_{i2} = CAGR d'*ouput_lait* calculé lorsque l'exploitation i a 6 ans ou plus.

Résultats

Le test a été mené pour différents y et chaque résultat est présenté dans le tableau 1.3.
En moyenne 5% des tests pour lesquels l'hypothèse H_0 est *réellement* vérifiée vont rejeter celle-ci à tort, cette proportion est par définition égale au risque de première espèce que nous nous sommes fixés.

En regardant les différences obtenues dans la première partie du tableau 1.3, un premier constat est celui d'un niveau moyen de production plus fort après la cinquième année d'existence. En effet, en dépit de la diminution du prix moyen du lait (*prix_lait* dans le tableau), d'environ 8.80€ pour 1000L, les exploitations observées avant et après leur cinquième année dégagent pour l'atelier lait un chiffre d'affaires (*output_lait*) supérieur en moyenne de 13 000€ sur la seconde période. Cela n'est permis que grâce à une hausse significative de la quantité totale de lait produite (*qt_lait*), de 49 600 litres en moyenne.
Cet accroissement du niveau de production s'accompagne-t-il d'une augmentation des facteurs de production ? Concernant le facteur travail, l'hypothèse d'égalité des niveaux moyens d'UTH (*uth*) n'est pas rejetée. Une distinction est à opérer quant au facteur capital. Alors que les niveaux moyens d'immobilisations matérielles (*immo_mat_net*) ou de construction (*immo_const_net*) ne connaissent pas d'évolutions significatives, le cheptel, en taille (*nb_vaches*) et en valeur (*biens_vivants_net*) connaît lui bien un gain significatif. Couplé à une plus forte productivité annuelle des vaches (*lait_prod_vache*), d'environ 200L par unité, cela pourrait expliquer le gain constaté de niveau moyen de production laitière en seconde période.
Ces exploitations sont-elles pour autant moins performantes économiquement lors de leurs toutes premières années d'activité ? Ce n'est pas en tout cas que ce que suggère les niveaux moyens d'EBE (*ebe_corr*), pour lesquels l'hypothèse nulle n'est pas rejetée. Finalement, résultat plus rassurant que surprenant, l'endettement moyen (*tx_endettement*) semble diminuer avec l'âge de l'exploitation. Lors de leur installation, les exploitations ont en effet des besoins importants en capital et sont donc plus endettées que des exploitations plus avancées dans leur cycle de vie.

Intéressons-nous désormais non plus au niveau moyen mais à la "dynamique" des outcomes y sur chaque période. Cette dynamique est retracée au travers des CAGR. La seconde partie du tableau 1.3 nous montre que pour les immobilisations nettes hors foncier (*immo*), le CAGR n'est pas significativement différent d'une période à une autre. On observe en revanche une plus forte intensification de la production après la cinquième année, avec un CAGR de l'ouput en moyenne supérieur de 5.34 points au niveau de toute l'exploitation (*output*), et en moyenne supérieur de 3.24 points pour l'atelier lait uniquement (*output_lait*).

Discussion

Si notre raisonnement par différenciation nous a permis faire fi de potentiels biais d'effets individuels fixes, cela n'est pas encore suffisant pour prétendre mesurer un effet causal de *pos_crea* sur nos outcomes y. Pour ces derniers, la théorie économique propose en effet bien des déterminants, qui ne sont pas seulement fixes mais aussi variables. Considérons comme outcome la différence d'immobilisations nettes moyennes avant et après seuil des cinq ans. Il est par exemple clair que les évolutions dans la conjoncture, la profitabilité de l'exploitation, l'âge de l'exploitant, ... sont autant de facteurs jouant un rôle important pour expliquer la différence obtenue. Voilà donc des variables, qui, si elles ne sont pas prises en compte, peuvent fausser nos comparaisons. C'est sur la considération d'une partie de ces déterminants qu'il va désormais s'agir de travailler.

6. Voir section 1.1 Quelle *derived variable*

TABLE 1.3 – Estimation and significance test for μ_D

	$\hat{\mu}_D$	$\hat{\sigma}_D$	z-value	95lower	95upper	p-value
Moyenne annuelle avant						
- moyenne annuelle après 5 ans pour :						
output	12320.62	39453.45	4.86	7349.83	17291.41	0
output_lait	13004.18	19388.29	10.43	10561.42	15446.93	0
prix_lait	-8.80	14.62	-9.36	-10.64	-6.96	0
qt_lait	49603.43	61520.03	12.54	41852.45	57354.42	0
lait_prod_vache	215.11	650.12	5.15	133.20	297.02	0
sau	5.39	11.88	7.06	3.90	6.89	0
uth	0	0.40	-0.12	-0.05	0.05	0.90
immo_mat_net	297.95	35393.61	0.13	-4161.33	4757.24	0.90
immo_const_net	1678.53	58347.86	0.45	-5672.79	9029.85	0.66
biens_vivants_net	3400.87	9963.75	5.31	2145.52	4656.21	0
immo	5377.35	79404.34	1.05	-4626.90	15381.60	0.29
nb_vaches	5.39	7.76	10.80	4.41	6.36	0
tx_endettement	-4.81	13.47	-5.55	-6.50	-3.11	0
ebe_corr	685.32	17136.12	0.62	-1473.68	2844.32	0.53
CAGR avant - CAGR après 5 ans pour :						
output	5.34	12	6.92	3.83	6.85	0
output_lait	3.24	13.12	3.84	1.59	4.89	0
prix_lait	2.11	4.65	7.08	1.53	2.70	0
qt_lait	1.05	12.73	1.28	-0.55	2.65	0.20
lait_prod_vache	-0.52	9.45	-0.85	-1.71	0.67	0.39
immo	1.08	22.35	0.75	-1.74	3.89	0.45
biens_vivants_net	-1.47	36.65	-0.62	-6.08	3.15	0.53
nb_vaches	1.15	12.15	1.47	-0.38	2.68	0.14

1.2 Utiliser le matching pour une inférence causale

Afin d'affiner notre inférence concernant les effets de la position dans son cycle de vie d'une exploitation (toujours au sens nouvel installé/déjà installé), nous décidons d'adopter une approche multidimensionnelle. Comme discuté plus haut (voir section 1.1.2), nombreux sont les facteurs pouvant jouer favorablement ou défavorablement sur la décision d'investissement d'une exploitation. Il se peut que ces facteurs ne soient pas répartis aléatoirement entre nouveaux installés et exploitations plus anciennes. Le but de cette partie va donc être de prendre en compte cette potentielle endogénéité, à travers une méthode d'appariement sur le score de propension.
Le sens que l'on va désormais donner à l'effet "nouvel installé" va quelque peu différer de celui donné auparavant. En effet jusqu'ici une exploitation était considérée l'année t comme "nouvelle installée" si son âge ne dépassait pas 5 ans ; les observations post-cinquième année n'étaient alors pas considérées comme correspondantes à ce statut. Désormais, le statut de nouvel installé d'une exploitation est fixé une fois pour toute sur la période. Autrement dit, sont considérées comme nouvelles installées les exploitations créées entre 2005 et 2015. Ce choix est discutable car près de 40% de celles-ci ont déjà soit 4 ou 5 ans lorsqu'elles sont observées pour la première fois. Il résulte néanmoins d'un compromis, du doute que l'on peut avoir quant à la validité d'un seuil commun de 5 ans pour définir la période d'installation de toutes les exploitations. [7]

1.2.1 Méthodologie

Pour comprendre les fondements de l'estimation par appariement, il nous faut faire un retour rapide sur le cadre de Rubin.

7. Nous reviendrons sur ce point dans la seconde partie du rapport, dédiée à l'analyse des trajectoires.

Cadre de Rubin

Dans les années 1970, Rubin propose un cadre formel adapté à l'évaluation statistique d'un traitement. Soit donc un traitement T dont on cherche à mesurer l'effet sur une variable d'outcome Y (par exemple, le revenu). On dispose d'individus, divisés en deux groupes : traités ($T_i = 1$) et non-traités ($T_i = 0$). Une notion qui est essentielle est celle de *revenu potentiel*. Pour chaque individu, on suppose en effet l'existence de deux revenus théoriques notés Y_{0i} et Y_{1i}. Y_{0i} correspond au revenu potentiel de l'individu i si celui-ci n'est pas traité, Y_{1i} si l'est. Par nature, seul l'un de ces deux revenus est observé ; d'où la qualification de *revenus potentiels*.
En considérant ces revenus potentiels, on peut mesurer l'effet du traitement de deux façon :

— via l'*Average Treatment effect on the Treated* (ATT) : $\Delta_{ATT} = \mathbb{E}(Y_{1i} - Y_{0i}|T_i = 1)$
— via l'*Average Treatment Effect* (ATE) : $\Delta_{ATE} = \mathbb{E}(Y_{1i} - Y_{0i})$

Nous allons pour notre part nous concentrer sur l'ATT. Notons Y_i le revenu observé : $Y_i = Y_{0i}|T_i = 0$ et $Y_i = Y_{1i}|T_i = 1$. Lorsque $Y_0 \perp\!\!\!\perp T$, on peut simplement estimer Δ_{ATT} par $\hat{\Delta}_{ATT} = \overline{Y}_{1\bullet} - \overline{Y}_{0\bullet}$ (avec $\overline{Y}_{1\bullet}$ la moyenne empirique chez les traités, $\overline{Y}_{0\bullet}$ celle chez les non-traités). En pratique, cette hypothèse n'est cependant jamais réalisée (hormis dans le cadre des expériences aléatoires) ; on dit qu'il y a alors un *effet de sélection*[8].
Il existe plusieurs méthodes pour réduire au mieux ce biais de sélection, parmi lesquelles on trouve la régression linéaire ou les méthodes de sélection sur observables.
Dans celles-ci l'hypothèse suffisante pour estimer un effet causal du traitement sur les traités est que, conditionnellement aux observables X, le fait d'être traité est indépendant du revenu potentiel sans traitement c'est-à-dire que l'on ait $Y_{0i} \perp\!\!\!\perp T_i|X_i$. Sous cette hypothèse, on peut espérer notamment construire le meilleur contre-factuel possible $\hat{g}(x_i)$ de chaque individu traité, et par là fournir une estimation de l'Δ_{ATT} de la forme (avec E_1 qui correspond à l'ensemble des individus traités et $N_1 = card(E_1)$) :

$$\hat{\Delta}_{ATT} = \frac{1}{N_1} \sum_{i \in E_1} (Y_{1i} - \hat{g}(x_i)) \tag{1.1}$$

Appariement sur le score de propension (PSM)

Un avantage des méthodes d'appariement sur la régression linéaire est que cette dernière impose une spécification linéaire qui n'est pas forcément adaptée. En revanche, une faiblesse des méthodes d'appariement "traditionnelles" réside dans la sensibilité des résultats à la fois à la métrique choisie et au nombre de variables de contrôle utilisées.
Une alternative au problème de dimensionnalité est proposée par Rosenbaum & Rubin (1983)[4] au travers de l'appariement sur le score de propension (PSM[9]). Notons $p(X) = \mathbb{P}(T = 1|X)$. Le développement de cette méthode repose sur une propriété montrée par Rosenbaum & Rubin (1983)[4], à savoir :

$$Y_0 \perp\!\!\!\perp T|X \Rightarrow Y_0 \perp\!\!\!\perp T|p(X)$$

En pratique, cela signifie que pour trouver un bon contre-factuel aux individus traités, il n'est plus nécessaire de regarder directement l'ensemble des observables X mais uniquement leur score de propension $p(X)$. Plus deux individus auront des scores proches, plus ils seront jugés semblables sur la base des observables.
La limite de cette méthode reste que $p(X)$ est très souvent inconnu et qu'il faut donc en fournir une estimation. Par conséquent, la qualité de l'estimation finale de Δ_{ATT} est fonction de la qualité de l'inférence préalable sur $p(X)$. Nous spécifierons $p(X)$ sous une forme logit, forme très largement utilisée pour ce genre d'étude.
Lorsque $\hat{p}(X)$ est calculé, plusieurs estimateurs de Δ_{ATT} sont possibles. Le plus simple est l'estimateur par appariement au plus proche voisin ; dans le cadre du PSM, celui-ci est donné par :

$$\hat{\Delta}_{ATT-PSM(NN)} = \frac{1}{N_1} \sum_{i \in E_1} (Y_{1i} - \hat{Y}_{0i}) \tag{1.2}$$

8. De façon générale, la terminologie utilisée est largement empruntée à l'épidémiologie, discipline dont ces réflexions sont principalement issues.
9. PSM : *Propensity Score Matching*

où : — $\hat{Y}_{0i} = Y_{j^*}$ avec $j^* = \underset{j \in E_0}{\text{argmin}} |\hat{p}(X_j) - \hat{p}(X_j)|$ (avec ou sans remise)

Des estimateurs quasi-équivalents consistent à ne pas utiliser qu'un seul mais plusieurs individus pour estimer le contre-factuel $\hat{Y}_{0i}$. C'est le cas des estimateurs dits par *radius* ou par *kernel*. Dans le premier cas, on définit un $h' > 0$ tel que l'on considère tous les individus non traités vérifiant $|\hat{p}(X_i - \hat{p}(X_j)| < h'$. Le second fournit une estimation non-paramétrique par noyau du type $\hat{Y}_{0i} = \frac{\sum_{j \in E_0} K(\hat{p}(X_i) - \hat{p}(X_j)).Y_j}{\sum_{j \in E_0} K(\hat{p}(X_i) - \hat{p}(X_j))}$.
Par manque de temps lors de ce stage pour réaliser des analyses de robustesse selon l'estimateur choisi, l'ATT ne sera estimé que suivant l'appariement par le plus proche voisin, donc suivant l'équation 1.2. Juger de la significativité de l'effet du traitement est une chose complexe avec les méthodes par appariement. En effet, la variance de notre estimateur est fonction à la fois de la variance de l'effet réel, de la variance de $\hat{p}(X)$, et de l'algorithme de matching utilisé. Une estimation de la variance par bootstrap est généralement utilisée comme alternative à ce problème.
Comme nous nous restreignons à une estimation par appariement au plus proche voisin (sans remise), nous avons jugé pertinent de procéder directement à un test de comparaison de moyenne sur échantillon apparié (avec comme paires chacun des couples traité/non-traité matchés). Si la significativité testée ne peut être interprétée comme celle de l'effet, elle n'en reste pas moins un indicateur. Cette méthode simple de test de comparaison de moyennes est souvent utilisée dans la littérature.

Choix des covariables

Le choix des covariables, c'est-à-dire des variables de contrôle X, est loin d'être anodin. Rappelons que pour que les estimateurs de l'ATT présentés plus haut soient pertinents, il faut que l'indépendance conditionnelle $Y_0 \perp\!\!\!\perp T|X$ soit vérifiée.
Nous serions intuitivement tenté d'inclure dans X le maximum de caractéristiques observables. Ce faisant, nous serions cependant confrontés aux risques suivants :

- inclure dans X des variables endogènes. C'est notamment le cas de variables mesurées après traitement et dont on peut penser qu'elles ont été influencées par celui-ci. Il peut s'agir de variables de résultats ou de pratiques agricoles, comme le niveau de pesticides ou d'antibiotiques, le chargement animal, etc.
- inclure dans X des variables dont le pouvoir explicatif sur la probabilité d'être traité est trop fort. Cela peut venir en effet enfreindre une hypothèse dont nous n'avions pas fait mention jusque là : l'hypothèse de support commun. Sa violation se traduit techniquement par le fait que si les individus traités et non-traités sont trop séparés au sens du score, l'appariement est rendu impossible. C'est un risque limité pour nos données.

1.2.2 Application

Dans les applications qui suivent, nous définissons le traitement T comme le fait d'être une exploitation nouvellement installée (au sens où nous l'entendons dans cette partie ; voir plus haut 1.2). Les outcomes Y étudiées ici seront uniquement des CAGR notés τ, calculés pour différentes variables et pour chaque exploitation sur toute sa période d'observation. Nous nous intéressons spécifiquement à l'effet du statut "nouvel installé" sur la croissance observée des exploitations, en termes de niveau de production (au travers des CAGR d'*output*, d'*output_lait* et de *qt_lait*) et en termes de facteurs de production (au travers des CAGR de *nb_vaches*, *tot_immo_net*, *immo* et *biens_vivants_net*). Nous décidons d'inclure successivement en tant que variables de contrôle (i) des variables liées à la période d'observation et au statut juridique, (ii) des variables liées au prix du lait, (iii) des variables liées à la structure des exploitations et (iv) des variables liées au niveau de capital initial.

Only controling for observation-period and status differences (Model 1)

Dans un premier temps, les variables incluses dans notre modèle de sélection (modèle par lequel on définit $\rho(X)$) sont uniquement les suivantes : les années de première et de dernière observation ainsi que le statut juridique (encore au sens individuel/sociétaire).
L'idée est de regarder si l'inclusion de ces quelques facteurs permet déjà d'apporter des nuances quant aux conclusions apportées par la simple comparaison de la moyenne des traités ($\overline{\tau}_{1\bullet}$) et celle des non-traités ($\overline{\tau}_{0\bullet}$) sur l'échantillon complet.

Including milk price (Model 2)

Aux variables de contrôle présentées précédemment nous ajoutons ensuite le prix moyen du lait. Pour chaque exploitation, elle correspond à la moyenne de ses prix de vente annuels (*prix_lait*) pendant qu'elle est observée. On espère ainsi corriger un potentiel biais lié à la conjoncture du lait.

Adding structural covariates (Model 3)

Nous ajoutons ensuite les variables de contrôle suivantes : le niveau initial d'UTH (*uth*), de SAU (*sau*) et de spécialisation laitière (*specialisation_marge*). [10] Ces variables, si elles sont prises à leurs niveaux annuels, peuvent être endogènes (par exemple, le niveau d'UTH peut être modifié si le nouvel installé utilise plus ou moins de main d'œuvre salariée). Nous faisons l'hypothèse que prendre les niveaux initiaux réduit ce problème d'endogénéité.

Adding initial capital size (Model 4)

Finalement, l'idée est de raisonner également à des niveaux comparables de capital initial. L'inclusion d'une telle variable peut cependant induire un problème d'endogénéité (voir section 1.2.1).

Les résultats des différents appariements sont présentés dans le tableau 1.4. Comme nos estimations sont soumises à la fois à la convergence d'algorithmes d'estimation par maximum de vraisemblance (pour $\hat{p}(X)$), mais aussi à l'ordre par lequel les individus sont matchés, il convient d'en "tester" la sensibilité. À défaut de fournir une complète analyse de la robustesse de nos estimations, nous présentons en annexe B.0.2 un tableau obtenu selon le même algorithme que 1.4.

Le tableau 1.4 présente en premier lieu les résultats obtenus par le simple test de comparaison de la moyenne entre le groupe des traités et le groupe des non-traités (colonne 'ttest'). Dans ce cas, pour toutes les variables d'origine considérées (*output*, *output_lait*, *qt_lait*, ...), on remarque une plus forte croissance chez les exploitations que nous considérons comme nouvellement installées.
Un premier appariement (Model 1) permet d'apporter une nuance quant au résultat précédent. En effet, lorsque les traités sont "matchés" au plus proche voisin avec un non-traité observé environ sur la même période et avec le même statut juridique, les différences de croissance sont moins claires. Les traités présentent certes toujours une croissance significativement plus forte en matière d'*output* et de *biens_vivants_nets* (colonne 'M1-pair.ttest'). Cette différence positive en faveur des traités est cependant moins évidente pour ce qui est d'*output_lait* et de *qt_lait* ; elle l'est d'autant moins pour *nb_vaches*, *tot_immo_net* et *immo*.
Pour (Model 2), là aussi les ATT estimés sont tous positifs (colonne 'M2-ATT'). Là aussi, sur les individus matchés, les différences de croissance en faveur des traités ne semblent significatives qu'au niveau d'*output* et de *biens_vivants_nets* (colonne 'M2-pair.ttest').
Lorsque nous incluons des covariables de type structurel (Model 3), les ATT estimés (colonne 'M3-ATT') sont tous supérieurs à ceux obtenus jusque là. Par exemple, alors que l'ATT estimé de *qt_lait* est de 1.12 point suite au second appariement (Model 2), il est de 1.73 point suite au troisième (Model 3). C'est dans (Model 3) que les différences entre individus traités et non-traités matchés sont globalement les plus marquées (colonne M3-pair.ttest).
Avec (Model 4), notre but est de comparer la croissance des traités et des non-traités à niveaux égaux d'immobilisations nettes initiales (hors foncier). Les résultats obtenus dans ce cas nous montrent, d'abord pour *immo*, que la croissance de ce capital pour les exploitations traitées et celle pour les exploitations non-traitées matchées ne sont pas significativement différentes (colonne 'M4-pair.ttest). En revanche, les résultats de ce matching indique que le traitement a bien un effet significativement positif sur la croissance d'*output*, d'*output_lait*, de *qt_lait* et de *biens_vivants_net*.

L'ATT estimé prend un sens à chaque fois différent selon les variables de contrôle que l'on considère. Tester différents appariements permet de comprendre les mécanismes à l'œuvre quant à l'origine des différences entre les nouveaux installés et les autres exploitations. Si elle devrait être complétée d'un diagnostic, notamment sur la qualité de l'appariement, notre analyse fait globalement apparaître les résultats suivant :

10. Par niveau initial on entend le niveau observé pour la première fois au niveau de l'exploitation.

— lorsque nous prenons en compte la totalité de l'échantillon (sans appariement), les nouveaux installés apparaissent comme particulièrement dynamiques par rapport au reste. Ils montrent un niveau de croissance moyen (CAGR) plus élevé en moyenne, et ce dans tous les domaines étudiés : valeur de production totale (*output*), valeur de la production laitière (*output_lait*), montant des immobilisations nettes (*tot_immo_net*, *immo*, *biens_vivants_net*)

— ces divergences s'amenuisent dès lors que l'on tente de raisonner à période d'observation et statut juridique équivalents (Model 1). Ne subsistent dans ce cas qu'un plus fort dynamisme en termes de niveaux de production (*output*, voire *output_lait*) et de valeur du cheptel (*biens_vivants_net*). Ce constat tend à perdurer au travers des différents modèles de sélection utilisées ensuite, excepté pour Model 3 où les différences de moyennes sont particulièrement prononcées pour presque tous les CAGR (voir colonne 'M3-pair.ttest').

TABLE 1.4 – Propensity score matching estimation
Note de lecture : 'CM' : Matched Control Group Mean, 'TM' : Matched Treated Group Mean, 'ttest' : p-value from mean difference t-test, 'pair.ttest' : p-value from paired t-test.

	$\overline{\tau}_{0\bullet}$	$\overline{\tau}_{1\bullet}$	$\overline{\tau}_{1\bullet} - \overline{\tau}_{0\bullet}$	ttest	M1-CM	M1-TM	$\hat{\Delta}_{ATT}$	M1-pair.ttest
CAGR for :								
output	-1.559	0.802	2.361	0	-0.302	0.823	1.125	0.008
output_lait	-0.388	2.678	3.066	0	1.273	2.732	1.460	0.064
qt_lait	0.688	3.362	2.674	0	2.034	3.418	1.383	0.076
nb_vaches	0.863	2.516	1.653	0	2.048	2.556	0.508	0.363
tot_immo_net	-0.928	0.406	1.335	0.014	0.038	0.423	0.385	0.539
immo	-1.891	0.194	2.085	0.001	-0.671	0.214	0.885	0.224
biens_vivants_net	-3.260	1.539	4.799	0	-1.285	1.597	2.882	0.001

	M2-CM	M2-TM	$\hat{\Delta}_{ATT}$	M2-pair.ttest	M3-CM	M3-TM	$\hat{\Delta}_{ATT}$	M3-pair.ttest
output	-0.348	0.823	1.172	0.010	-0.567	0.823	1.390	0.002
output_lait	1.572	2.732	1.160	0.137	0.823	2.732	1.909	0.009
qt_lait	2.290	3.418	1.128	0.149	1.682	3.418	1.736	0.018
nb_vaches	2.181	2.556	0.375	0.513	1.694	2.556	0.862	0.066
tot_immo_net	-0.099	0.423	0.522	0.381	-0.476	0.423	0.900	0.155
immo	-0.699	0.214	0.913	0.198	-1.050	0.214	1.265	0.078
biens_vivants_net	-1.420	1.597	3.017	0.001	-1.503	1.597	3.100	0

	M4-CM	M4-TM	$\hat{\Delta}_{ATT}$	M4-pair.ttest
output	-0.600	0.823	1.423	0.003
output_lait	0.831	2.732	1.902	0.017
qt_lait	1.758	3.418	1.659	0.037
nb_vaches	1.765	2.556	0.791	0.167
tot_immo_net	-0.137	0.423	0.560	0.393
immo	-0.634	0.214	0.849	0.257
biens_vivants_net	-1.217	1.597	2.814	0.001

Nous l'avons vu dans ce chapitre, raisonner dans un cadre statique permet de faire appel à de nombreuses analyses statistiques classiques puisque l'indépendance de nos observations y est vérifiée. Nous n'avons utilisé qu'une partie de celles-ci, pour répondre à une question particulière. Lorsque l'on s'intéresse à des trajectoires, se ramener à un cadre statique présente cependant un inconvénient majeur. Cela revient en effet à mettre de côté une grande partie de l'information dont nous disposons, puisque nous ignorons alors l'information portée par la période entre le premier et dernier moments d'observation (voir formule du CAGR).
Pour prendre en compte cette information "intermédiaire" jusque-là ignorée, il nous faut reconsidérer la totalité de notre panel. Nous allons désormais voir quelles sont les approches adaptées à de telles données.

Chapitre 2

Analyse des trajectoires

2.1 Quelle(s) méthode(s) utiliser ?

L'analyse de données répétées en général

Confronté à des données répétées, il est possible d'opter parmi plusieurs choix de modélisation pour mener à bien son analyse : séries temporelles, modèles de panel, modèles multiniveaux, modèles mixtes, ... Si la frontière entre ces modèles s'avère poreuse, et que les pratiques diffèrent suivant les disciplines, le choix de l'analyste va surtout dépendre (i) de la nature de la répétition des données, (ii) de la question à laquelle il souhaite répondre.
Les séries temporelles sont utilisées dans un cadre bien particulier, celui où un nombre réduit d'indicateurs sont suivis sur une période relativement longue, pour lesquels l'intérêt porte plutôt sur la mise en évidence de dynamiques tendancielles et/ou cycliques. Si les données correspondent plutôt à des mesures répétées dans un temps relativement court sur un grand nombre d'individus, **et** que l'objet de l'analyse porte sur la mesure d'un effet causal, il convient là plutôt de recourir à des modèles de panel. Une modélisation du type multiniveaux est en revanche privilégiée lorsque les mesures ont été répétées au sein de groupes (résultats scolaires d'élèves au sein d'une classe, patients traités dans un même hôpital, ...) **et** lorsque l'intérêt porte éventuellement sur la différenciation de ces groupes.
Évidemment, et comme souligné dans ce qui suit, des nombreuses exceptions résistent à cette brève typologie.

Intérêt des "Growth Curve Models"

Nos données correspondent à des mesures répétées au niveau d'un grand nombre d'exploitations, suivis annuellement sur au plus 10 ans. Pourquoi alors avoir emprunté dans la suite une modélisation de type multi-niveaux et non de type panel par exemple ?
Pour le comprendre, revenons-en au sujet qu'il s'agit ici de traiter : l'analyse des trajectoires d'investissement. Plus précisément, l'intérêt n'est pas d'estimer ce qui serait un *effet causal* du temps. Il est plutôt d'apprécier les évolutions intraindividuelles et de prendre la mesure des différences d'évolutions entre les individus ; autrement dit d'apprécier la forme et l'hétérogénéité des trajectoires individuelles.
C'est justement à ce type de problème, à l'origine surtout soulevé en sciences comportementales et en biostatistique, auxquelles sont dédiées des approches relativement récentes, sans réelle unification théorique mais rassemblées sous le nom de *Growth Curve Modeling*[1].
Le principe général de ces méthodes est assez simple. Elles consistent à spécifier une forme fonctionnelle commune à toutes les trajectoires, puis à étudier la variabilité des paramètres associés à cette forme, ceux-ci étant supposés différent d'un individu à un autre. Résumons cette idée via une simple équation, où on considère la j-ème observation de l'exploitation i, avec y un output quelconque et $farm\_age$ l'âge de l'exploitation : $y_{ij} = \pi_{0i} + \pi_{1i} farm\_age_{ij}$. Cette relation pose bien que (i) le lien entre l'âge d'une exploitation $farm\_age_{ij}$ et son niveau d'outcome y_{ij} est linéaire, (ii) le niveau d'outcome inital π_{0i} :*initial status* et la pente π_{1i} :*rate of change* diffèrent selon l'exploitation considérée.
Deux approches vont dans ce sens : celle des *Multilevel Models for Change* (hierarchical models, mixed-effects models, linear mixed model...) et celle des *Latent Growth Models* (ou latent curve models).

1. ou *Growth Curve Analysis*

Notre démarche

Parce qu'elle fait appel à des notions généralement plus familières, nous emprunterons dans un premier temps l'approche par modélisation multiniveaux (*Multilevel Models for Change)*. Nous verrons comment celle-ci, dans sa formalisation, inclut distinctement les deux niveaux d'analyse qui nous intéressent : le niveau intra et interindividuel. Nous présenterons ensuite l'intérêt et la souplesse de cette méthode au travers de diverses applications, en lien avec nos questionnements sur les trajectoires de capital net (hors foncier) (*immo*) des exploitations nouvellement installées.
Après avoir constaté dans un premier temps une très forte hétérogénéité des trajectoires, nous décidons de changer légèrement de point de vue. Le but de notre seconde partie sera plutôt d'identifier des *pools* de trajectoires, c'est-à-dire de dégager de l'ensemble des trajectoires, et de façon purement empirique, des sous-groupes d'exploitations suivant à peu près la même direction. De nombreuses techniques semblent avoir été développées dans ce sens depuis les années 2000, parmi lesquelles les plus citées sont basées sur les *Growth Mixture Models* (mélange entre *finite mixture model* et *growth curve models*). Il s'agira pour nous de présenter un cas simple d'un tel modèle puis d'en illustrer l'intérêt sur un exemple.

2.2 Un premier pas dans l'analyse des trajectoires

2.2.1 Le modèle de référence

Dans le cadre d'une modélisation multi-niveaux, le modèle de base des *(Linear) Growth Curve Models* est communément appelé *uncondtionnal growth model*. En reprenant les notations de Singer & Willet (2003)[6], le modèle s'exprime ainsi :

$$\begin{cases} y_{ij} = \pi_{0i} + \pi_{1i} farm\_age_{ij} + \epsilon_{ij} & \text{(Level-1)} \\ \pi_{0i} = \gamma_{00} + \zeta_{0i} & \text{(Level-2.a)} \\ \pi_{1i} = \gamma_{10} + \zeta_{1i} & \text{(Level-2.b)} \end{cases}$$

$$\text{where : } (i) \quad \epsilon_{ij} \quad \text{i.i.d} \quad \mathcal{N}(0,\, \sigma_\epsilon^2)$$
$$(ii) \begin{pmatrix} \zeta_{0i} \\ \zeta_{1i} \end{pmatrix} \quad \text{i.i.d} \quad \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} \right)$$

Si nous sommes déjà familiers avec l'équation de "niveau 1" (Level-1) (voir section 2.1), la spécification complète du modèle laisse apercevoir deux équations supplémentaire, dites de "niveau 2" (Level-2.a et Level-2.b). Celles-ci expriment l'hypothèse posée sur (π_{0i},π_{1i}), vu comme un vecteur gaussien centré sur (γ_{00},γ_{10}). Elles expliquent ici la différence de trajectoires entre deux individus comme le résultat d'un aléa. Bien qu'admise, l'hétérogénéité inter-individuelle des trajectoires n'est pas du tout expliquée dans ce modèle, d'où l'emploi du terme *unconditionnal*. Évidemment un des intérêts d'une telle modélisation réside dans la facilité avec laquelle il est ensuite possible d'introduire des termes explicatifs, aussi bien dans le niveau 1 que dans le niveau 2 ; c'est-à-dire des variables influant directement le niveau y_{ij} (time-varying predictors) ou influant les paramètres π_{0i} et π_{1i} (time-invariant predictors). Avant d'entrer dans ces considérations, appréhendons les trajectoires de la façon la plus simple, de façon visuelle.

2.2.2 Quelques graphiques préalables

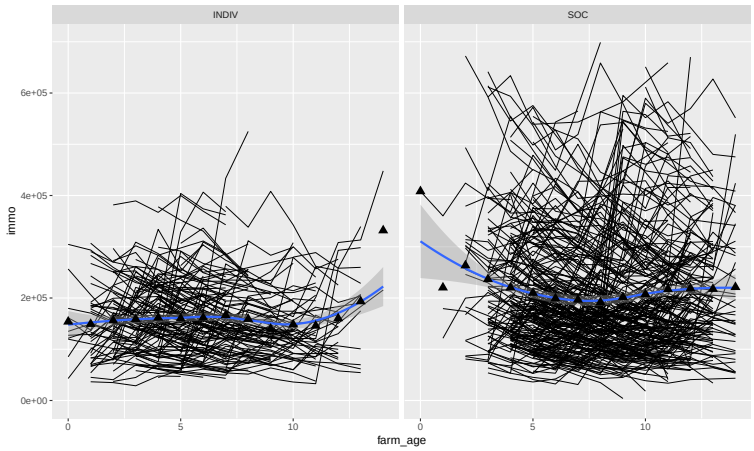
Toute tentative de modélisation des trajectoires requiert au préalable une analyse exploratoire de ces dernières. Ce travail en amont, qui sera ici uniquement graphique, est particulièrement important car il permet de façon très simple :

1. de deviner la spécification la plus adaptée pour modéliser l'ensemble des trajectoires. Sont-elles par exemple plutôt linéaires ou quadratiques, continues ou discontinues ?
2. de rendre compte de l'hétérogénéité inter-exploitations des trajectoires, d'un point de vue du niveau initial de capital (π_{0i}), et de son évolution (π_{1i})
3. de juger graphiquement de l'influence de certaines variables, en tant que *time-varying predictor* ou en tant que *time-invariant predictor*

Nous présentons figure 2.1 la trajectoire d'*immo* des exploitations créées entre 2005 et 2014. Les trajectoires représentées correspondent à une interpolation des valeurs annuelles au niveau de chaque exploitation.
Nous choisissons ici d'exhiber séparément les exploitations individuelles des exploitations sociétaires. Outre le gain de visibilité obtenu (lorsque les trajectoires sont trop nombreuses, il est parfois proposé de ne représenter qu'un sous-échantillon aléatoire d'individus), cela permet d'identifier de potentielles différences de trajectoires entre les deux groupes. On peut évidemment étendre ce raisonnement à d'autres caractéristiques, ce qui est fait en annexe C.0.1.
Pour ce qui est des trajectoires obtenues sur la figure 2.1, plusieurs faits sont à noter. La grande majorité des exploitations nouvellement installées ne sont pas observées durant leur première année d'existence (cas où $farm\_age_{ij}$ vaut 0). Elles sont par ailleurs observées sur des durées variables puisque certaines apparaissent après 2005 et d'autres ne sont pas observées jusqu'en 2010.

Nous remarquons d'abord une plus forte hétérogénéité des niveaux d'immobilisations pour les exploitations sociétaires (à droite), avec par ailleurs une trajectoire en moyenne plus élevée (les trajectoires en bleu sont des estimations non-paramètriques de $immo_{ij}$ en fonction de $farm\_age_{ij}$).

FIGURE 2.1 – Net assets (excluding land) trajectories over farm age



La forme fonctionnelle de la régression est toujours un choix de spécification délicat. Illustrons ce point en nous focalisant sur le graphique de gauche (pour les exploitations individuelles). On peut y voir grossièrement deux types de trajectoires. Un premier type rassemble des trajectoires plutôt linéaires dans le temps, avec un niveau d'immobilisations légèrement décroissant voire constant. Certaines trajectoires connaissent en revanche des ruptures nettes qui révèlent un investissement conséquent pour les années concernées. Cette typologie visuelle illustre la difficulté à trouver une spécification adaptée à toutes les trajectoires. La présence de ces ruptures annonce déjà le faible pouvoir explicatif du modèle de base (*unconditionnal growth model*), qu'il s'agira peut-être d'améliorer en introduisant des variables explicatives appropriées.

2.2.3 Spécification de modèles

Nous allons dans cette partie considérer différents modèles pour expliquer les trajectoires observées, ils incluent tour à tour des variables prédictives de niveau 1 ou de niveau 2. Ces modèles sont volontairement simplistes : nous n'envisagerons ici pas de trajectoire du type polynomiale par exemple. Encore une fois, il s'agit moins pour nous d'expliquer parfaitement les trajectoires que d'illustrer l'interprétation de certains paramètres
Le modèle élémentaire des modèles de croissance est un modèle sans croissance (Model 1). Il sert en effet de référence statistique pour juger de l'apport de $farm\_age_{ij}$ en tant que variable explicative de $immo_{ij}$. Son rôle est aussi de rendre compte de la part de variabilité totale de y_{ij} associée à la variabilité intraindividuelle. Dit autrement, il permet de quantifier le niveau de dépendance intraindividuelle de nos observations (dépendance qui motive rappelons-le toutes les méthodes utilisées dans

nos travaux). Cette mesure se fait à travers le rapport de corrélation $\rho = \frac{\sigma_0^2}{\sigma_0^2+\sigma_\epsilon^2}$ (où σ_ϵ^2 représente la variabilité résiduelle de y, σ_0^2 la variabilité inter-individuelle de y ; voir Model 1 ci-dessous).

Model 1 - Unconditional Means Model

On suppose les ϵ_{ij} i.i.d $\mathcal{N}(0,\, \sigma_\epsilon^2)$ et ζ_{0i} i.i.d $\mathcal{N}(0,\, \sigma_0^2)$ (ϵ_{ij} et ζ_{0i} non-corrélés) dans :

$$\begin{cases} immo_{ij} = \pi_{0i} + \epsilon_{ij} & \text{(Level-1)} \\ \pi_{0i} = \gamma_{00} + \zeta_{0i} & \text{(Level-2.a)} \end{cases}$$

NOTE

Pour chacun des modèles qui vont suivre (Model 2-3-4-5) on suppose :
(i) ϵ_{ij} i.i.d $\mathcal{N}(0,\, \sigma_\epsilon^2)$
(ii) $\begin{pmatrix}\zeta_{0i}\\ \zeta_{1i}\end{pmatrix}$ i.i.d $\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_0^2 & \sigma_{01}\\ \sigma_{01} & \sigma_1^2\end{pmatrix}\right)$
(iii) $\epsilon \perp\!\!\!\perp \begin{pmatrix}\zeta_0\\ \zeta_1\end{pmatrix}$

Model 2 - Unconditional Growth Model : voir section 2.2.1

Model 3 - Conditional Growth Model (a)

$$\begin{cases} immo_{ij} = \pi_{0i} + \pi_{1i} farm\_age_{ij} + \beta dummy\_context_{ij} + \epsilon_{ij} & \text{(Level-1)} \\ \pi_{0i} = \gamma_{00} + \zeta_{0i} & \text{(Level-2.a)} \\ \pi_{1i} = \gamma_{10} + \zeta_{1i} & \text{(Level-2.b)} \end{cases}$$

Dans Model 3, nous incluons en tant que variable prédictive de niveau 1 (*time-varying predictor*) une variable appelée *dummy_context* (vaut 1 lorsque la j-ème observation sur i est prise entre 2006 et 2008, 0 sinon). En ajoutant cette variable, notre but est de tester la présence d'éventuelles ruptures (notamment des désinvestissements) liées au contexte de crise du lait sur cette période. L'effet de ce contexte, β, est supposé constant à travers les exploitations.

Model 4 - Conditional Growth Model (b)

$$\begin{cases} immo_{ij} = \pi_{0i} + \pi_{1i} farm\_age_{ij} + \beta dummy\_context_{ij} + \epsilon_{ij} & \text{(Level-1)} \\ \pi_{0i} = \gamma_{00} + \gamma_{01} \mathbb{1}_{SOC_i} + \zeta_{0i} & \text{(Level-2.a)} \\ \pi_{1i} = \gamma_{10} + \gamma_{11} \mathbb{1}_{SOC_i} + \zeta_{1i} & \text{(Level-2.b)} \end{cases}$$

Le statut juridique a-t-il une influence sur la trajectoire de capital net des exploitations nouvellement installées ? Pour répondre à cette question, Model 4 inclut 1_{SOC} (vaut 1 si le statut est sociétaire, 0 si individuel) en tant que variable explicative des paramètres π_0 et π_1. En procédant ainsi, on pourra tester si ce facteur a une influence (i) uniquement sur le niveau initial de capital (π_0), (ii) uniquement sur la pente de la trajectoire de capital (π_1) ou (iii) à la fois sur π_0 et sur π_1.

Model 5 - Conditional Growth Model (c)

$$\begin{cases} immo_{ij} = \pi_{0i} + \pi_{1i} farm\_age_{ij} + \beta dummy\_context_{ij} + \epsilon_{ij} & \text{(Level-1)} \\ \pi_{0i} = \gamma_{00} + \gamma_{01} \mathbb{1}_{SOC_i} + \gamma_{02} \mathbb{1}_{IFA\_Q2Q3_i} + \gamma_{03} \mathbb{1}_{IFA\_Q4_i} + \zeta_{0i} & \text{(Level-2.a)} \\ \pi_{1i} = \gamma_{10} + \gamma_{12} \mathbb{1}_{IFA\_Q2Q3_i} + \gamma_{13} \mathbb{1}_{IFA\_Q4_i} + \zeta_{1i} & \text{(Level-2.b)} \end{cases}$$

Après inspection des résultats obtenus pour Model 4 (voir section 2.2.4), nous décidons de retirer le statut juridique des facteurs pouvant expliquer la pente de la trajectoire π_1 (c'est-à-dire de Level-2.b). Nous nous intéressons ensuite à l'effet de l'âge de l'exploitant sur la trajectoire d'une exploitation nouvellement installée. Pour de multiples raisons (liées à la motivation, au dynamisme, à la prise

de risque etc.), on peut en effet penser qu'un exploitant plus jeune au moment de son installation aura une propension plus forte à investir qu'un exploitant beaucoup plus âgé. Si cet effet est avéré, la pente π_{1i} sera plus élevé pour l'exploitation avec le plus jeune exploitant. La variable 1_{IFA_Q2Q3} est une indicatrice d'appartenance aux second et troisième quartiles de la variable $init\_farmer\_age$. Autrement dit, $1_{IFA\_Q2Q3_i}$=1 si, au premier instant d'observation de l'exploitation i, l'exploitant (ou le plus vieux des associés pour une forme sociétaire) appartient à la tranche d'âge des [30,50] ans (dans notre échantillon, 25% des exploitants sont âgés de moins de 30 ans au moment où ils sont observés pour la première fois, 25% d'entre-eux ont plus de 50 ans). De la même façon $1_{IFA\_Q4_i}$=1 si l'exploitant (ou le plus vieux des associés si forme sociétaire) de i a plus de 60 ans lorsqu'il est observé pour la première fois.

2.2.4 Interprétation des résultats

Les modèles sont estimés par maximisation de la log-vraisemblance restreinte (*restricted log-likelihood*). [2]
Le tableau 2.1 présente les paramètres estimés ainsi que des indicateurs de comparaison de modèles (AIC, BIC, Log Likelihood).
Commençons par le modèle sans croissance (Model 1). Dans ce modèle, $\hat{\gamma}_{00}$=195 000€ correspond au niveau moyen des immobilisations nettes (hors foncier) observées sur la période (les $immo_{ij}$) pour notre échantillon. Environ 74% des déviations à cette moyenne sont attribuables à des différences entre les individus, cette proportion est en effet égale à $\hat{\rho} = \frac{\hat{\sigma}^2_\epsilon}{\hat{\sigma}^2_\epsilon + \hat{\sigma}^2_0}$. Cela signifie que seulement 26% de la variabilité totale est attribuable à de la variabilité intra-individuelle (σ_ϵ) ; indiquant donc une forte dépendance des observations au sein d'une même exploitation.
Dans le cadre d'un modèle avec croissance linéaire (Model 2), $\hat{\gamma}_{00}$ et $\hat{\gamma}_{10}$ estiment respectivement le niveau moyen de départ (lorsque $farm\_age$=0) et le niveau moyen de la pente des trajectoires. On trouve qu'en moyenne, le niveau initial (lorsque $farm\_age_{ij} = 0$) d'immobilisations net (hors foncier) des exploitations est d'environ 181 000€($\hat{\gamma}_{00}$), avec une pente moyenne d'environ 1668€($\hat{\gamma}_{10}$), jugée non significative. Dans l'*uncondtionnal growth model*, $\hat{\sigma_0}$ et $\hat{\sigma_1}$ sont des indicateurs de la dispersion individuelle autour de cette trajectoire moyenne. Leur signification est légèrement différente lorsqu'on ajoute des variables prédictives de niveau 2 (voir Model 4 et 5), puisque σ^2_0 représente alors la variance inexpliquée de π_0 ; σ^2_1 la variance inexpliquée de π_1.
Par rapport à Model 2, Model 3 permet aux trajectoires linéaires de connaître des ruptures momentannées liées à la crise de 2007 (que l'on capte avec $dummy\_context_{ij}$). Au vu du tableau, ces années de crise ont un effet marginal jugé significativement négatif sur les niveaux $y_{ij} = immo_{ij}$, d'une valeur estimée de $\hat{\beta} = -6720$€.
L'estimation de Model 4 suggère que le statut sociétaire a un impact assez fort sur le niveau de départ des trajectoires. La mesure de cet effet, qui est donné par $\hat{\gamma}_{01}$, indique ici qu'une exploitation sociétaire possède en moyenne 57 988€ de plus lorsqu'elle s'installe qu'une exploitation individuelle. Il n'y a en revanche pas d'effet significatif du statut sociétaire sur la pente, γ_{11} n'est par conséquent pas conservé pour Model 5.
Dans Model 5, nous testons, au travers de variables catégorielles l'effet de l'âge de l'exploitant, à la fois sur le niveau initial et la pente des trajectoires. Tel que nous avons spécifié le modèle, les paramètres γ_{02}, γ_{03}, γ_{12} et γ_{22} s'interprètent comme des déviations au niveau initial (pour les deux premiers) et à la pente (pour les deux derniers) de la catégorie de référence : les exploitations dont l'âge de l'exploitant (ou du plus vieil associé) est inférieur 30 ans ($\mathbb{1}_{IFA\_Q4_i} = \mathbb{1}_{IFA\_Q4_i} = 0$). Ces déviations sont-elles importantes au vue des résultats obtenus ? Pour une exploitation, être géré par un exploitant qui a moins de 30 ans ou qui a entre 30 et 50 ans ne semble rien changer quant au niveau initial et à la pente de la trajectoire (γ_{02} et γ_{12} non-significatifs). Entre les exploitations avec exploitant "âgé" ($1_{IFA\_Q4_i} = 1$) et les exploitations avec exploitant "jeune" (la référence), il semble exister une différence significative dans la pente (γ_{13} significatif) mais pas dans le niveau initial (γ_{03} non-significatif).
En guise d'illustration, les trajectoires individuelles et les trajectoires moyennes estimées de chaque groupe (exploitations individuelles avec exploitant de moins de 30 ans, sociétaires avec exploitant de moins de 30 ans, individuelles avec exploitant entre 30 et 50 ans, ...) sont représentées sur la figure C.2 en annexe C.0.2.

2. Voir https ://cran.r-project.org/web/packages/nlme pour une documentation sur le package R utilisé, notamment la commande *lme*.

Les explications fournies précédemment ne constituent évidemment pas une analyse et un diagnostic complets des modèles estimés. Pour que les conclusions apportées par ces derniers soient suffisamment défendables, il faudrait s'assurer notamment de la relative tenabilité de nos hypothèses de linéarité, de normalité et d'homoscedasticité. La façon d'y procéder dans les modèles multi-niveaux ne diffère que très peu de celle des modèles de régression linéaires classiques.

TABLE 2.1 – Regression table from Model 1 to Model 5

	Model 1	Model 2	Model 3	Model 4	Model 5
Initial status					
γ_{00}	194180.65***	181612.66***	185842.25***	149505.60***	158892.92***
	(5539.28)	(7562.54)	(7644.47)	(12340.71)	(18051.32)
β			−6720.46***	−6855.23***	−6747.64***
			(1806.05)	(1807.01)	(1807.64)
γ_{01}				57988.42***	64943.01***
				(15414.85)	(11303.77)
γ_{02}					−12552.35
					(20174.07)
γ_{03}					−41413.76
					(29445.55)
Rate of change					
γ_{10}		1668.92	1360.34	2390.67	4838.71*
		(899.01)	(899.40)	(1538.62)	(2171.60)
γ_{11}				−1743.31	
				(1889.78)	
γ_{12}					−3648.65
					(2406.50)
γ_{13}					−8715.89*
					(3450.95)
AIC	74641.87	73405.20	73376.56	73323.26	73243.34
BIC	74659.85	73441.14	73418.49	73377.16	73315.20
Log Likelihood	-37317.94	-36696.60	-36681.28	-36652.63	-36609.67
Num. obs.	2954	2954	2954	2954	2954
Num. groups	363	363	363	363	363

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 2.2 – Random Effect Estimated Standard Deviation

	Model 1	Model 2	Model 3	Model 4	Model 5
σ_ϵ	61293.58	42111.27	42019.70	42025.52	42042.14
σ_0	103210.47	136343.15	136312.03	133461.69	133264.27
σ_1		15813.87	15748.89	15733.85	15583.28
$\rho_{01} = \frac{\sigma_{01}}{\sigma_0 \sigma_1}$		-0.685	-0.685	-0.689	-0.707

L'hétérogénéité observée et quantifiée des trajectoires de capital sur les premières années d'activité est très forte. Nous n'avons réussi qu'à capter une très légère partie de celle-ci en introduisant *soit* des variables expliquant directement y_{ij} (niveau 1), *soit* des variables expliquant les paramètres (π_{0i},π_{1i}) (niveau 2). À nouveau, on ne peut cependant être surpris de ce résultat, tant les comportements d'investissement dépendent d'une multitude de facteurs. Rappelons ici que ce stage s'inscrit dans un ensemble de travaux de recherche qui traitent beaucoup plus largement de cette thématique.Ces raisons nous poussent à changer de point de vue quant à l'analyse de nos trajectoires.

2.3 Détection de groupes homogènes de trajectoires

Les travaux effectués dans la partie précédente ont révélé une grande hétérogénéité dans les évolutions d'immobilisations des nouveaux installés. Au lieu d'expliquer ces comportements sous un angle éco-

nomique, il va s'agir pour nous d'adopter un point de vue beaucoup plus empirique, visant d'abord à révéler derrière cette hétérogénéité d'éventuels "groupes" de trajectoires.
C'est justement dans ce sens qu'ont été développés les *Growth Mixture Models* (Muthén&Muthén ;2000)[9]. Ces modèles sont plus exactement une adaptation des *finite mixture models* à l'analyse des trajectoires. Les *finite mixture models* s'utilisent lorsque l'on pense que notre échantillon total se décompose en plusieurs sous-échantillons, générés chacun suivant un processus aléatoire différent. Le modèle qui est censé avoir généré l'échantillon total est donc un mélange (*mixture*) d'un nombre de fini (*finite*) de sous-modèles, et à chaque sous-modèle correspond un groupe d'observations qui aura été généré selon celui-ci. Lorsque l'on s'intéresse à des évolutions (*Growth Mixture Model*), ces sous-modèles prennent la forme de *growth curve models*, desquels sont donc séparément générés des sous-échantillons de trajectoires. Ces groupes ou sous-échantillons sont dits "cachés" (on parle dans la littérature de *latent class*) car qu'ils ne sont pas définis a priori mais délimités de façon exploratoire.

Dans un premier temps nous verrons comment se formalise un cas particulier de *Growth Mixture Model*, où l'on considère des spécifications uniquement linéaires pour nos trajectoires ("linéaire" ici au sens général du terme, c'est-à-dire comprenant des formes polynomiales par exemple), et qui ne nécessite pas de faire appel à la théorie des *Latent Growth Models*.
Via un tel *Latent Class Linear Mixed Model* (LCLMM), nous verrons ensuite s'il est possible d'identifier des groupes de trajectoires sur nos données. Là aussi, notre but ne sera pas de fournir une analyse complète appliquée à une problématique ; simplement d'illustrer l'intérêt de cette approche à travers un rapide exemple.

2.3.1 Latent Class Linear Mixed Model

On va supposer notre échantillon d'exploitations divisé en G sous-populations latentes, notées g (g=1, ..., G). Pour toute exploitation i, c_i est aléatoire et sa réalisation désigne l'unique groupe auquel cette exploitation appartient ($c_i \in \{1, ..., G\}$). La probabilité que i appartienne au groupe g est notée $\pi_{ig} = P(c_i = g)$.
On peut également poser un modèle plus élaboré qui décrit la probabilité d'appartenance à un groupe selon certaines caractéristiques individuelles X_{1i}, par exemple de la forme (ici suivant une spécification logistique) :

$$\pi_{ig} = P(c_i = g|X_{1i}) = \frac{e^{\xi_{0g}+X'_{1i}\xi_{1g}}}{\sum_{l=1}^{G} e^{\xi_{0l}+X'_{1i}\xi_{1l}}}$$

La trajectoire des exploitations, conditionnellement à l'appartenance de celles-ci au groupe g, sont ensuite modélisées suivant un *(Linear) Growth Model* (voir section 2.2). Lorsque qu'aucune variable de niveau 1 ou de niveau 2 n'est envisagée (c'est-à-dire avec un *unconditionnal growth model*), les G sous-modèles de croissance peuvent s'écrire simplement :

$$immo_{ij|c_i=g} = \pi_{0ig} + \pi_{1ig} farm\_age_{ij} + \epsilon_{ij} \tag{2.1}$$

$$\begin{aligned} \text{avec : } (i) \quad & \epsilon_{ij} \quad \text{i.i.d} \quad \mathcal{N}(0, \sigma_\epsilon^2) \\ (ii) \quad & (\pi_{0ig}, \pi_{1ig})' \sim \mathcal{N}(\mu_g, B_g) \end{aligned}$$

Nous remarquons dans l'écriture 2.1 que les trajectoires des exploitations appartenant au groupe g oscillent autour d'une trajectoire moyenne donnée par μ_g (bidimensionnel), spécifique à ce groupe. La dispersion autour de cette trajectoire moyenne est donnée par B_g, qui peut elle aussi être spécifique à g ; le plus souvent on pose *soit* $B_g = B$ (la dispersion est la même dans tous les groupes), *soit* $B_g = w_g^2 B$.

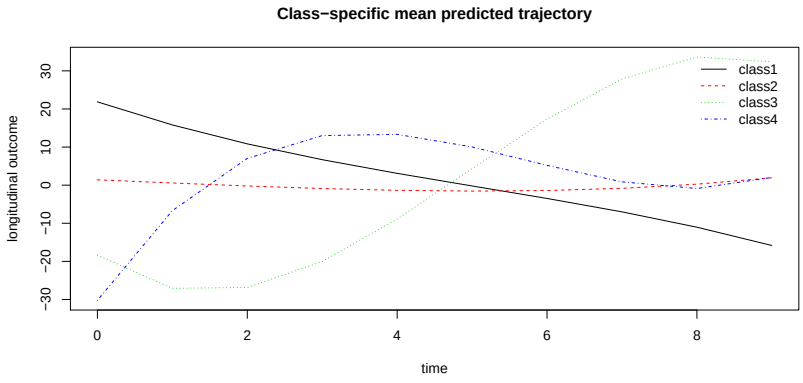
Comme son nom l'indique, le LCLMM peut donc être vu comme un mélange de G sous-modèles de croissance "classiques" (comme étudiés dans la section 2.2). Le fait qu'une exploitation suive la direction moyenne donnée par un sous-modèle plutôt que celle d'un autre est aléatoire. On peut néanmoins faire dépendre la probabilité d'appartenir à un groupe de certaines caractéristiques intrinsèques (que nous avons notées X_{1i}).

Lorsque nous supposons qu'un tel modèle a généré nos données, il convient d'estimer les paramètres qui y sont associés. Dans le package R que nous avons utilisé, cela est réalisé au travers d'un algorithme EM de la log-vraisemblance (voir Proust&Lima ;2010)[10].

2.3.2 Illustration sur l'échantillon cylindré

Nous allons désormais illustrer le genre de résultats qu'il est possible d'obtenir avec cette technique, cette fois-ci en ne nous intéressant pas aux exploitations nouvellement installées mais à l'échantillon cylindré. Cela permet d'écarter d'éventuels problèmes de convergence liés à un manque de données. La notion de trajectoire va donc prendre un sens légèrement différent de celui que nous nous étions fixé jusque là. Par "trajectoire" nous n'entendrons désormais plus la liaison individuelle liant l'âge de l'exploitation et à un indicateur mais celle qui lie ce dernier au "temps", c'est-à-dire la variable t_{ij} avec $t_{ij} \in \{0, ..., 9\}$ (indices des années 2005 à 2014).
Pour des raisons purement numériques, nous décidons par ailleurs de ne pas étudier directement les trajectoires d'immobilisations nettes (hors terrain) ($immo$) mais celles données par centrage et réduction de ces niveaux ; c'est à dire les trajectoires de la variable $norm\_immo_{ij} = \frac{immo_{ij} - \overline{immo}_{..}}{\hat{\sigma}(immo_{ij})}$. Après inspection graphique des trajectoires obtenues par interpolation des valeurs annuelles (voir figure C.3 en annexe C), nous envisageons 3 à 4 sous-populations (il est apparu après estimation que le modèle avec $G = 4$ présentait un AIC et un BIC plus faibles), avec pour chaque groupe des trajectoires cubiques (linéaires d'ordre 3).
Les trajectoires moyennes estimées des 4 sous-groupes obtenus sont présentées dans la figure 2.2. En affectant les exploitations à un groupe en suivant la règle $\hat{c}_i = \underset{g \in \{1,..,4\}}{\operatorname{argmax}} \hat{\pi}_{ig}$ nous obtenons que, sur un total de 596 exploitations, 2.35% de celles-ci appartiennent à 'class-1', 88.16% à 'class-2', 4.53% à 'class-3' et 4.36% à 'class-4'.

FIGURE 2.2 – Mean estimated trajectory of each group



Nous avons vu à travers un exemple que la méthode des LCLMM permet d'identifier des comportements "types". Suivant les besoins de l'étude, il peut être intéressant de tester des spécifications différentes, un nombre de groupes G plus/moins élevé. Les sous-modèles de croissance estimés dans le LCLMM s'interprètent comme des *Latent Curve Model*. La possibilité d'inclure des variables prédictives (X_{1i}) à l'appartenance des groupes peut également être un aspect exploité pour de futures analyses sur ce sujet (ce que nous n'avons pas fait pour notre exemple).
Lorsqu'un des modèles LCLMM testés est jugé satisfaisant, une étape qui suit peut suivre dans l'analyse consiste à comparer les sous-populations obtenues sur la base de différents critères d'intérêt, de façon à trouver des pistes d'explications quant aux différences obtenues. Cette étape n'est pas réalisée ici par manque de temps lors de ce stage

Conclusion

Les principales questions auxquelles nous nous sommes chargés de répondre durant ce stage ont été les suivantes :

— Peut-on affirmer que les cinq années qui suivent la création d'une exploitation agricole représentent une période particulière pour celle-ci ?

— Dès lors que l'on s'intéresse à des trajectoires, quelles sont méthodes statistiques les plus adaptées ? Comment se formalisent-elles et comment en interpréter les résultats ?

Dans un premier temps de ce rapport, nous avons avons répondu à la première question dans un cadre statique simple, en nous ramenant à deux indicateurs résumant la situation des exploitations sur la période : leur moyenne temporelle et leur CAGR. Une ANOVA sur les CAGR a révélé des différences de niveaux moyens entre nouveaux entrants et "anciens", et avec ou non le statut sociétaire. Le groupe présentant le CAGR moyen le plus élevé, de 1.8 points environ, rassemblait les nouveaux entrants de type "individuel". En nous concentrant ensuite sur le sous-échantillon des exploitations observées avant et après leur cinquième d'existence, il a semblé que les niveaux moyens et les croissances de capital net étaient les mêmes avant et après ce seuil, ce constat s'inversait pour la valeur de la production par exemple. Parce que le CAGR observé peut s'expliquer par d'autres facteurs, nous proposons finalement plusieurs méthodes d'estimation par appariement. Les effets estimés obtenus s'interprètent alors différemment selon le modèle d'appariement retenu.

La deuxième partie du rapport s'intéresse plus largement aux trajectoires de capital net (hors foncier). Celle-ci invite d'abord à faire un point sur l'ensemble des approches qu'il est possible d'emprunter lorsque nous sommes confrontés à des données répétées. Elle tente de situer les *Growth Curve Models* par rapport à cet ensemble. Dans ce sens, nous avons vu (i) qu'ils s'utilisent lorsque l'on s'intéresse aux différences interindividuelles de trajectoires, (ii) qu'ils peuvent simplement s'exprimer sous la forme d'un modèle multi-niveaux. Nous avons ensuite utiliser cette méthode pour analyser la trajectoire des nouveaux entrants. Les quelques spécifications que nous avons testées n'ont pas suffit à appréhender leur très forte hétérogénéité. Il nous a alors fallu introduire les *Growth Mixture Models*. Cette approche suggère que, derrière cette forte dispersion, se cachent en réalité des sous-groupes relativement homogènes de trajectoires. Dans leur formalisation, ces modèles sont en fait un mélange de plusieurs quasi-*Growth Curve Models* qui modélisent chacun la trajectoire d'un sous-groupe. Appliqué au panel cylindré, et avec un nombre de groupes fixé au préalable, cette méthode exploratoire a permis d'identifier quatre trajectoires "types" pour les exploitations observées entre 2005 et 2014.

En termes de limites à porter à nos travaux, je pense que la question des données manquantes est particulièrement importante. Nous l'avons ignorée pour des raisons de temps, la plupart du temps en ne considérant que les exploitations observées un nombre minimum de fois (voire en ne prenant que le panel cylindré). L'hypothèse implicite que nous avons alors faite, et qui mérite d'être questionnée, est celle d'une indépendance de l'inobservation aux phénomènes que nous avons étudiés.

Si nous avions des extensions à apporter à notre analyse, elles seraient focalisées sur la détection de sous-groupes de trajectoires. Nous n'avons fait qu'illustrer cette méthode sur un exemple. Je trouve cependant que cette approche est particulièrement attrayante pour détecter d'éventuels déterminants à certaines formes de trajectoires, qu'elle mérite d'être approfondie sur ces données.

Bibliographie

[1] Boehlje M., « The Entry-Growth-Exit Processes in Agriculture », *Southern Journal of Agricultural Economics*, 1973.

[2] Diggle P. et al., *Analysis of Longitudinal Data*, Oxford Statistical Science Series, 2002.

[3] Givord P., « Méthodes économétriques pour l'évaluation de politiques publiques », *Economie & prévision*, 2014.

[4] Rosenbaum P.,Rubin D., « The Centrale Role of the Propensity Score in Observational Studies for Causal Effects », *Biometrika*, 1983.

[5] Caliendo M., Kopeining S., « Some Practical Guidance for the Implementation of Propensity Score Matching », *Journal of Economic Survey*, 2005.

[6] Singer Judith D.,Willet John.B, *Applied Longitudinal Data Analysis*, Oxford University Press, 2003.

[7] Curran P.J., Obeidat K., Losardo D., « Twelve Frequently Asked Questions About Growth Curve Modeling », *Journal of cognition and development*, 2010.

[8] Hox J., Stoel R., « Multilevel and SEM Approaches to Growth Curve Modelling », *in* Everitt & Howell, *Encyclopedia of Statistics in Behavioral Science*, 2005.

[9] Muthén B, Muthén L.K., « Integrating person-centered and variable-centered analysis : Growth mixture modeling with latent trajectory classes », *Alcoholism : Clinical and Experimental Research*, 2000.

[10] Proust-Lima C., Philipps V., Liquet B., « Estimation of Extended Mixed Models Using Latent Classes and Latent Processes : The R Package lcmm », *Journal of Statistical Software*, 2017.

Annexe A

Données et traitements préalables

A.0.1 Présentation des bases disponibles

Les données dont nous disposons pour cette étude sont issues de documents comptables réalisés chaque année entre 2005 et 2014 par Cerfrance. Elles concernent au total 4 482 exploitations agricoles, toutes localisées en Ille-et-Vilaine. Il se peut que des exploitations, soit disparaissent, soit apparaissent d'une année à l'autre dans le panel. Ainsi, pour 4 482 exploitations enregistrées dans la base, "seules" 28 672 observations sont récoltées durant les dix années de suivi (avec la définition, une observation = une exploitation prise une année donnée).

Une autre base recense différentes dates concernant les exploitations observées. Pour une exploitation individuelle sont fournies les dates de naissance de l'exploitant ainsi que celle de son entreprise. Lorsqu'il s'agit d'une forme sociétaire (exploitation collaborative), sont fournies, en plus de la date de création de l'entreprise, les dates de naissance de chacun des associés.

A.0.2 Mise en forme et "nettoyage"

Dans le panel, certaines observations correspondaient à des exercices de moins ou de plus de 12 mois, expliquant une grande partie des doublons observés sur le couple (*numadh*,*annees*) (voir table A.1). Néanmoins, même en ne gardant que des observations avec une durée d'exercice égale 12 mois, des doublons étaient toujours présents. Cela correspondait à des décalages temporels entre la variable *annees* renseignée et les variables de dates de début et de fin d'exercice. Les observations concernées ont été retirées du panel.

De même nous avons extrait du panel les exploitations dont le statut juridique correspondait à des cas très particuliers, non considérés pour l'étude. Ainsi, n'ont été gardées que les exploitations présentant une des formes juridiques suivantes : forme individuelle, GAEC, EARL ou SCEA. [1]

Certaines observations présentaient des valeurs aberrantes au sens comptable puis statistique du termes. Il a donc s'agit, *soit* de retirer complètement l'observation concernée (notamment lorsque la valeur est caractéristique d'un comportement spécifique exclu de l'étude), *soit* de garder celle-ci en imposant toutefois une valeur manquante pour les variables présentant une valeur aberrante.

Ce sont au total 4 005 exploitations que nous conservons, associées à 24 094 observations.

A.0.3 Dictionnaire des variables

Modalités des variables catégorielles :

- *statut* : vaut "Indiv" si l'exploitation est individuelle (un seul gérant), vaut "Soc" si l'exploitation est cogérée (GAEC, SARL ou SCEA).
- *pos_crea* : vaut "creaSup" si l'exploitation a 6 ans plus, "creaInf" sinon.
- *cat* : vaut "IndivcreaSup", "IndivcreaInf", "SoccreaSup" ou "SoccreaInf"

1. Voir http ://chambres-agriculture.fr pour connaître les spécificités

Remarque sur *ebe_corr* **:**

Au niveau comptable, pour une exploitation individuelle, l'Excédent Brut d'Exploitation comprend la rémunération de l'exploitant alors que cette même donnée ne comprend pas la rémunération des associés pour les formes sociétaires ; d'où le besoin de correction.

TABLE A.1 – Dictionnaire des variables

Nom	Signification
numadh	Identifiant de l'exploitation
annees	Année d'observation
statut	Statut juridique
sau	Surface agricole utile (ha)
uth	Unité de travail humain (uth)
uth_ha	Quantité d'UTH par hectare de SAU (uth/ha)
tot_immo_net	Valeur totale nette des immobilisations avec foncier (eur.csts)
immo	Valeur totale nette des immobilisations hors foncier (eur.csts)
immo_const_net	... de construction (eur.csts)
immo_mat_net	... de matériel (eur.csts)
biens_vivants_net	... de biens vivants (eur.csts)
part_mais_sfp	Surface de maïs produits/Surface de fourrages produits (en %)
nb_vaches	Effectif moyen de vaches
chargement	Chargement à l'ha (ugb/ha)
qt_lait	Quantité totale de lait produite (litres)
lait_prod_vache	Quantité de lait produite par vache laitière (litres/vache)
prix_lait	Prix moyen de vente pour 1000l. (eur.csts)
output_lait	Valeur de la production laitière (eur.csts)
output	Valeur de la production totale (eur.csts)
specialisation_marge	Marge brute atelier lait/ Marge brute totale (en %) [2]
specialisation_prod	Valeur production lait/Valeur production totale (en %) [3]
tx_endettement	Montant dettes financières/Montant total des actifs (en %)
capital_prdty	Productivité du capital (en € de production/€ de capital)
ebe_corr	Valeur de l'EBE corrigé (eur.csts)
date_crea	Date de création de l'exploitation
date_last_stlmt	Date de dernière installation
pos_crea	Date de création > ou <= à 5 ans ?
cat	Croisement des variables *qualit* et *pos_crea*
farm_age	Âge de l'exploitation
init_farmer_age	Âge initial du plus vieil exploitant (de l'unique exploitant si ferme individuelle)
$\mathbb{1}_{IFA\_Q2Q3_i}$	Vaut 1 si l'exploitation appartient aux second ou troisième quartiles de $init\_farmer\_age$, 0 sinon
$\mathbb{1}_{IFA\_Q4_i}$	Vaut 1 si l'exploitation appartient au dernier quartile de $init\_farmer\_age$, 0 sinon

2. Le ratio est mis à 0 si (i) le ratio calculé est négatif, ou (ii) la marge brute totale est négative. Si le ratio calculé est supérieur à 1, *specialisation_marge* vaut 100%.

3. Pour 4 observations, le taux de spécialisation calculé est supérieur à 100%, il est dans ce cas majoré à ce seuil.

Annexe B

Compléments sur l'approche statique

B.0.1 Statistiques descriptives

TABLE B.1 – Summary statistics for the net assets (field assets not included)

Statistic		N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Beginning value	$Y_{i,1}$	2 847	190 514.10	139 656.40	5 410	96 698	243 578	1 897 123
Ending value	Y_{i,k_i}	2 847	194 980.80	163 876.60	591	92 173.5	248 126.5	1 900 891
Year difference	k_i	2 847	5.44	3.09	1	2	9	9
Absolute growth	$Y_{i,k_i} - Y_{i,1}$	2 847	4 466.75	106 524.70	−599 947	−40 604.5	33 027.5	1 264 312
Relative growth	$\frac{Y_{i,k_i} - Y_{i,1}}{Y_{i,1}}$	2 847	6.97	58.05	−98.62	−25.34	24.33	556.74
CAGR	τ_i	2 847	−1.47	14.14	−88.25	−7.89	4.16	117.68

B.0.2 PSM : une seconde procédure d'estimation

TABLE B.2 – Propensity score matching estimation
Note : 'CM' : Matched Control Group Mean, 'TM' : Matched Treated Group Mean, 'ttest' : p-value from mean difference t-test, 'pair.ttest' : p-value from paired t-test.

	M1-CM	M1-TM	M1-ATT	M1-pair.ttest	M2-CM	M2-TM	M2-ATT	M2-pair.ttest
output	-0.395	0.823	1.218	0.005	-0.348	0.823	1.172	0.010
output_lait	1.107	2.732	1.625	0.043	1.572	2.732	1.160	0.137
qt_lait	1.880	3.418	1.538	0.053	2.290	3.418	1.128	0.149
nb_vaches	1.696	2.556	0.860	0.144	2.181	2.556	0.375	0.513
tot_immo_net	-0.758	0.423	1.182	0.047	-0.099	0.423	0.522	0.381
immo	-1.756	0.214	1.970	0.006	-0.699	0.214	0.913	0.198
biens_vivants_net	-2.940	1.597	4.537	0	-1.420	1.597	3.017	0.001

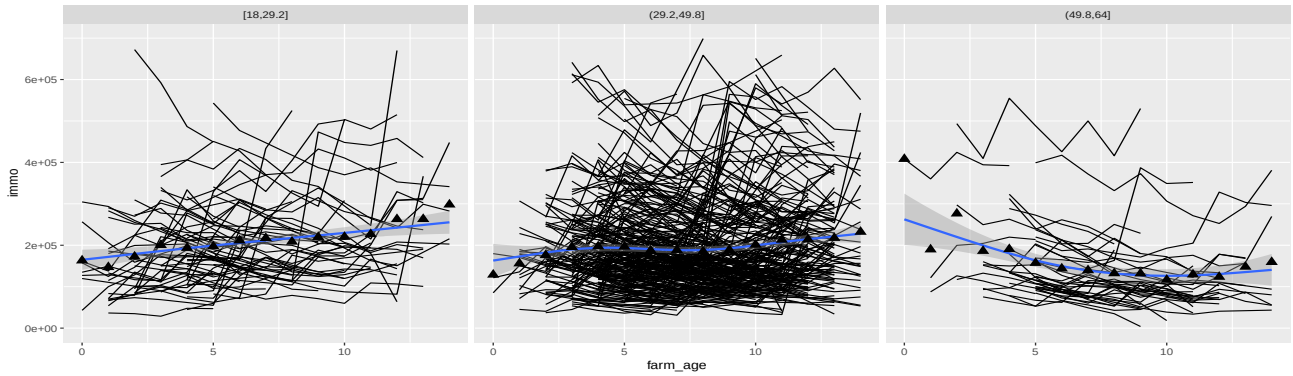
	M3-CM	M3-TM	M3-ATT	M3-pair.ttest	M4-CM	M4-TM	M4-ATT	M4-pair.ttest
CAGR for :								
output	-0.567	0.823	1.390	0.002	-0.600	0.823	1.423	0.003
output_lait	0.823	2.732	1.909	0.009	0.831	2.732	1.902	0.017
qt_lait	1.682	3.418	1.736	0.018	1.758	3.418	1.659	0.037
nb_vaches	1.694	2.556	0.862	0.066	1.765	2.556	0.791	0.167
tot_immo_net	-0.476	0.423	0.900	0.155	-0.137	0.423	0.560	0.393
immo	-1.050	0.214	1.265	0.078	-0.634	0.214	0.849	0.257
biens_vivants_net	-1.503	1.597	3.100	0	-1.217	1.597	2.814	0.001

Annexe C

Compléments sur l'analyse des trajectoires

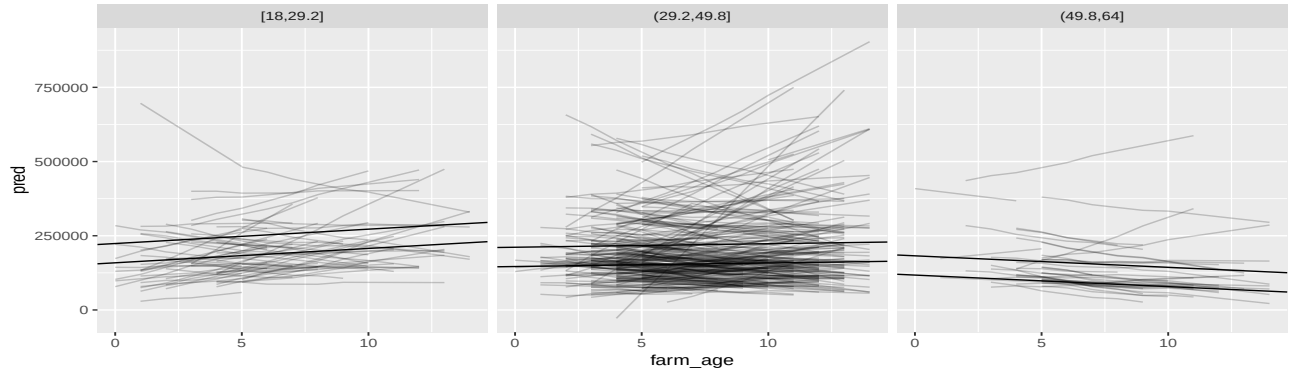
C.0.1 Trajectoires séparées selon l'âge de l'exploitant

FIGURE C.1 – Net assets observed trajectories for incoming farms
Note : Left panel displays the trajectories from farms belonging to $init\_farmer\_age$ first quartile. Middle panel - second and third quartile. Right panel - top quartile.



C.0.2 Trajectoires estimées de Model 5

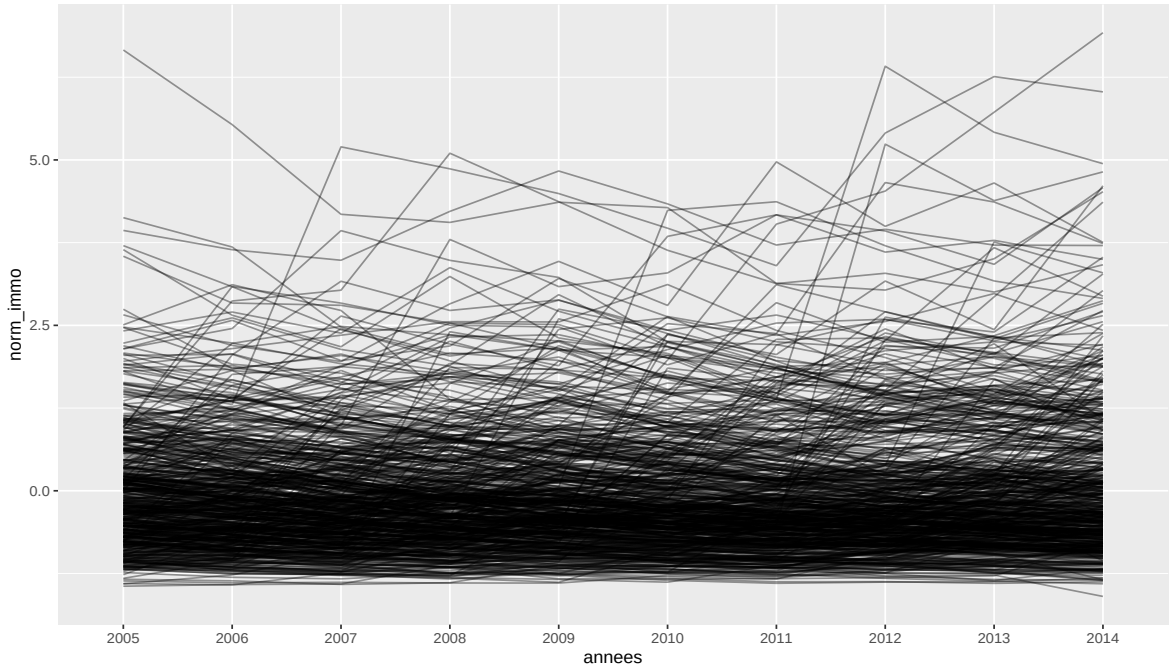
FIGURE C.2 – Individual and group fitted trajectories from Model 5
Note : Left panel displays the fitted trajectories from farms belonging to $init\_farmer\_age$ first quartile. Middle panel - second and third quartile. Right panel - top quartile.



C.0.3 Trajectoires sur l'échantillon cylindré

FIGURE C.3 – Standardized nets assets observed trajectories for the cylindric sub-sample (farms observed 10 times)

Note : Standardized values represent $norm\_immo_{ij} = \frac{immo_{ij} - \overline{immo}_{..}}{\hat{\sigma}(immo_{ij})}$. This transformation doesn't affect neither the order nor the form of the trajectories, only the magnitude of their scattering.



Annexe D

Accomplissements personnels

En termes d'accomplissements personnels, cette expérience professionnel a été particulièrement enrichissante. Ce stage a été pour ma part la première expérience professionnel réellement en lien avec mon futur travail de statisticien.

Les travaux qui m'ont été confiés m'ont d'abord permis de mettre en application mes acquis théoriques de première et de seconde année. J'ai spécialement apprécié la liberté qui m'a été accordée dans les choix méthodologiques pour traiter le sujet. Cela m'a en effet permis de prendre du recul par rapport à mes propres connaissances, mais aussi de les enrichir via de multiples recherches bibliographiques. Dans ce sens, j'ai trouvé les méthodes d'analyse de trajectoire (*Growth Curve Analysis*) très épanouissantes à étudier et à appliquer.

Enfin, cette expérience au sein d'une unité de recherche en économie agricole m'a permis d'appréhender à sa manière à la fois le monde de la recherche et celui de l'agriculture.