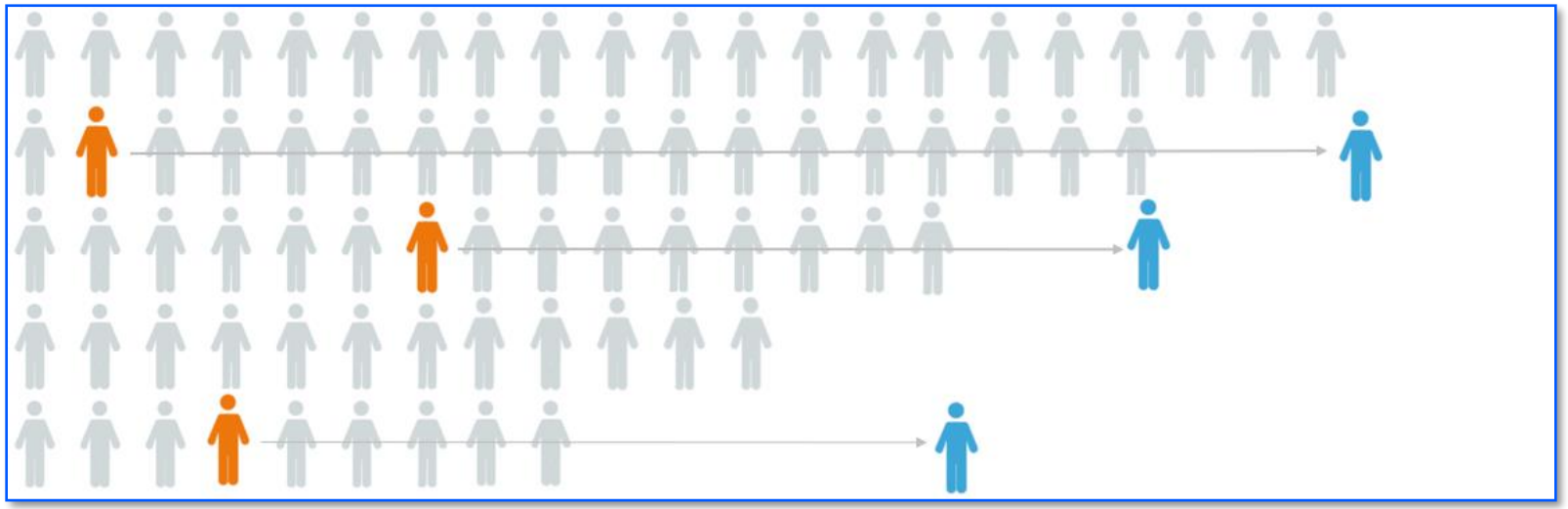
Des modèles agrégés aux modèles individuels : le provisionnement Non-Vie en pleine mutation

Franck Yombi
Direction Actuariat Groupe
Groupama

Eve Titon
R&D
Milliman



Introduction

- Depuis une dizaine d'années, de plus en plus de données individuelles riches en informations sur les sinistres sont collectées.
- En parallèle, de nombreux modèles ont été développés, permettant de tirer parti de ces données, à des fins explicatives ou prédictives.
- Dans le cadre du provisionnement, les méthodes individuelles pourraient permettre:
  - de pallier certains défauts des méthodes classiques agrégées
  - d'obtenir des estimations plus précises, sur des branches où le montant du sinistre dépend beaucoup des caractéristiques du sinistre

Contexte – partenariat entre Groupama et Milliman

RC Auto corporel sur des sinistres graves

- Plusieurs dizaines de millions de malis de graves ces dernières années (2018, 2019, 2020)
- Forte volatilité des dossiers graves au cours de leur développement
- Forte disparité de la sinistralité RC Auto corporel entre les différentes caisses régionales
- IFRS 17 et le suivi de la sinistralité à une maille plus fine

Plan de la présentation

1. **Exploration des données individuelles**
   - Présentation des variables exploitables
   - Périmètre de l'étude
   - Exemples d'analyses effectuées

2. **Implémentation de quelques modèles individuels**
   - Avantages et inconvénients des modèles testés
   - Résultats obtenus

3. **Enjeux et limites du passage d'une approche agrégée à une approche individuelle**
   - Aspects normatifs
   - Aspects opérationnels

4. **Perspectives et conclusion**

Exploration des données individuelles

Panorama des variables exploitables – cas de la RC corporelle

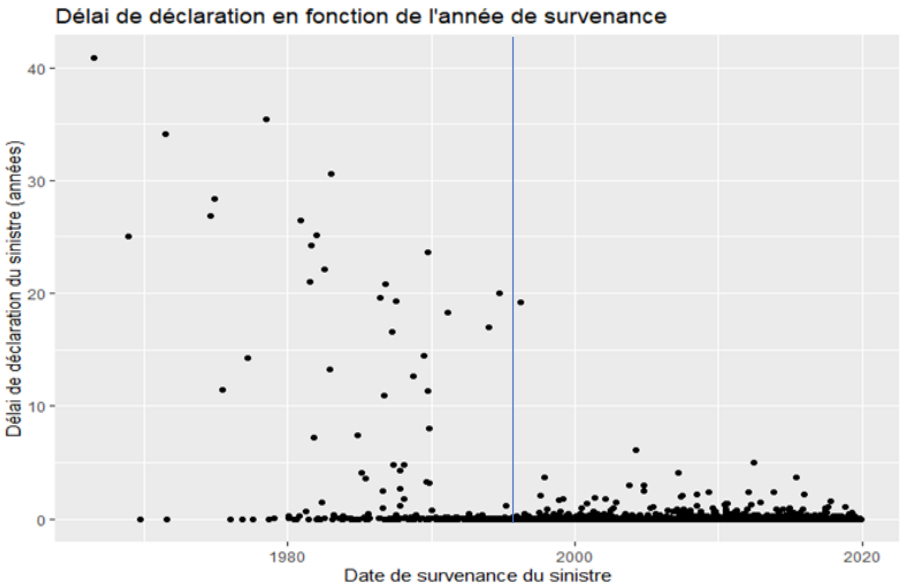
Variables concernant les provisions	Variables concernant le sinistre et l'assuré	Variables concernant la victime
Ensemble des variables qui concernent la provision effectuée par les gestionnaires sinistres, à chaque temps de vie du sinistre. Exemples: ▪ Montant de la provision initiale ▪ Révisions de provisions ▪ Caractéristiques du gestionnaires (âge, région, etc.) Ces variables permettent de faire des études sur: ▪ la fiabilité de la provision émise par les gestionnaires ▪ l'identification des facteurs qui influencent le provisionnement	Variables sur l'assuré, telles que: ▪ Age du véhicule ▪ Age du permis ▪ Age du conducteur ▪ Région, code postal Variables concernant le sinistre, telles que: ▪ Type d'événement ▪ Circonstances du sinistre ▪ Nombre de blessés	▪ Gravité des blessures ▪ Age de la victime ▪ Type de prise en charge de la victime ▪ Taux d'AIPP

Exploration des données individuelles

Périmètre de l'étude

➢ Restrictions de la base étudiée

- Conservation des sous-dossiers victimes graves uniquement (on exclut les sous-dossiers victimes non graves rattachés à un sinistre grave) afin d'éviter tout biais d'observation des victimes non graves.
- Restriction aux sinistres survenus à partir du 01/01/1996, pour deux raisons
  - On observe peu de sinistres survenus avant cette date. Par ailleurs, les sinistres survenus avant 1996 ont des délais de déclaration anormalement longs.
  - Cette restriction permet d'être en ligne avec le périmètre de calcul des triangles de projection utilisés lors des arrêtés comptables.



➢ Périmètre de modélisation

- Le cadre général du provisionnement des sinistres graves:

Provisions = RBNS + IBNR + IBNR graves

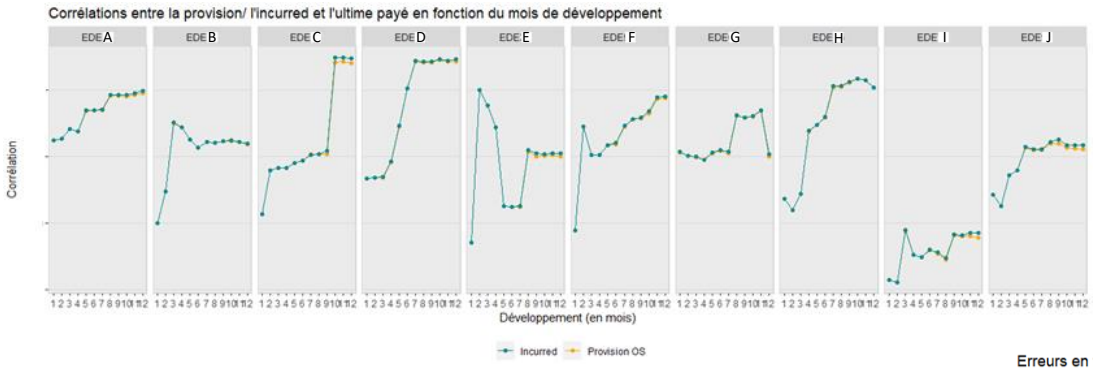
- L'estimation des IBNR et IBNR graves par une approche individuelle nécessite une étude spécifique

➔Dans la suite, on provisionne uniquement les RBNS.

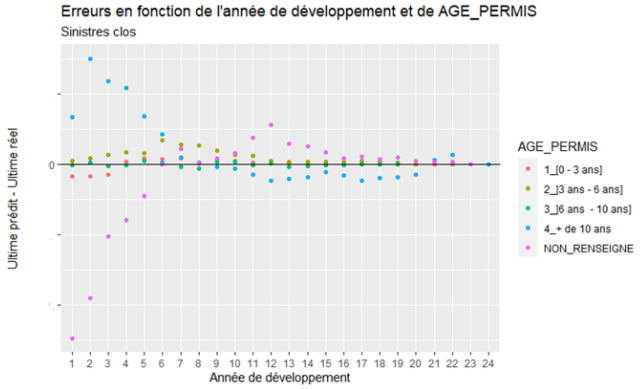
Exploration des données individuelles

Exemples d'études effectuées (1/2)

- **Etude de la fiabilité de la première provision émise par les gestionnaires sinistres** : corrélation entre la provision gestionnaire et l'ultime payé (étude effectuée sur l'ensemble des sinistres clos)
  - permet de déterminer à partir de quand la provision est fiable/est assez prédictive de l'ultime
  - permet d'identifier des axes d'amélioration de la procédure de provisionnement



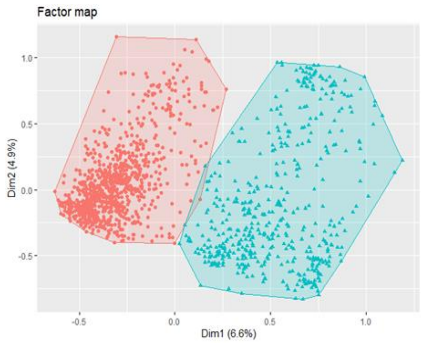
- **Etude de l'erreur de provisionnement en fonction de covariables et de l'année de développement du sinistre**
  - permet d'identifier, grâce à des statistiques univariées, sur quelle population les gestionnaires sinistres surestiment ou sous-estiment la provision à effectuer
  - possibilité d'identifier les facteurs expliquant l'erreur des gestionnaires (suppression des effets croisés entre les variables)

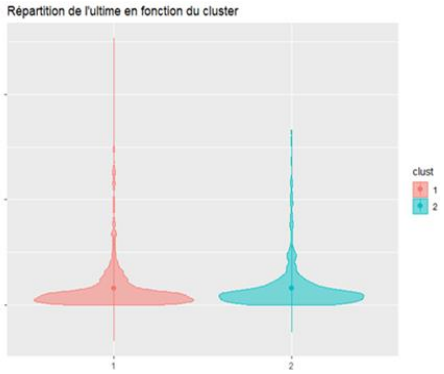


Exploration des données individuelles

Exemples d'études effectuées (2/2)

- **Segmentation a priori**





Cluster	Moyenne (Ultime)	SD(Ultime)	CV
1	1 033 395	1 631 254	1.58
2	1 028 221	1 368 574	1.33

- L'ACM montre dans un premier temps que les covariables présentes dans la base ne permettent pas d'expliquer la variabilité des caractéristiques des sinistres : le modèle explique seulement 12% de la variance.
- Les 2 clusters (déterminés grâce aux composantes principales de l'ACM) ne se distinguent pas par leur distribution d'ultimes. Ces classes se caractérisent par les covariables ci-dessous listées (par ordre d'importance les modalités qui définissent chacun des deux clusters).

Cluster 1	Cluster 2
AGE_PERMIS=AGE_PERMIS_4_+ de 10 ans	SEXE_CONDUCTEUR=SEXE_CONDUCTEUR_NR
METIER_PSO=AUTOMOBILE DE TOURISME	AGE_PERMIS=AGE_PERMIS_NON_RENSEIGNE
SEXE_CONDUCTEUR=SEXE_CONDUCTEUR_MASCULIN	METIER_PSO=TMA
CD_CSP=Salarié	CD_CSP=Personne Morale
SEXE_CONDUCTEUR=SEXE_CONDUCTEUR_FEMININ	METIER_PSO=FLOTTE
ZONE_CIRCONSCRIPTION=ACCIDENT_DANS_CIRCONSCRIPTION	AGE_VEH=AGE_VEH_NON_RENSEIGNE
CD_CSP=Retraité	ZONE_CIRCONSCRIPTION=ACCIDENT_HORS_CIRCONSCRIPTION
AGE_PERMIS=AGE_PERMIS_3_]6 ans - 10 ans]	caisse=EDE35
AGE_PERMIS=AGE_PERMIS_1_[0 - 3 ans]	METIER_PSO=AUTRES VEHICULES
AGE_PERMIS=AGE_PERMIS_2_]3 ans - 6 ans]	CD_CSP=Inactif
AGE_VEH=AGE_VEH_3_]7 ans - 10 ans]	Nb_victimes=Nb_victimes_1
Nb_victimes=Nb_victimes_3	LIB_TYPE_EVT_TERRAIN=AUTRES EVENEMENTS OU DOMMAGES
caisse=EDE31	METIER_PSO=AUTO ENTREPRISE(GARAGE)
delai_entre_declaration_et_seuil_500k=[0,1]	CD_CSP=Exploitant agri
Nb_victimes=Nb_victimes_>=4	USAGE_VEHICULE=USAGE_VEHICULE_AFFAIRE

Implémentation de quelques modèles individuels

Avantages et inconvénients des modèles testés

Modèle		Avantages	Inconvénients
Modèles non-paramétriques	Entraînement sur sinistres clos uniquement	• Intégration directe des covariables dans le modèle • Utilisation de l'ensemble des informations disponibles à chaque vision du sinistre • Intégration de la variable « passage en rentes »	• Biais de sélection (les sinistres clos peuvent avoir des développements plus courts et donc une tendance à avoir un Ultime plus faible) • Une proportion importante de sinistres est retirée de l'analyse, ce qui entraîne une perte d'informations
	Entraînement sur sinistres clos et non-clos développés par Chain Ladder	• Idem (covariables intégrées, utilisation de l'ensemble du dev des sinistres) • Permet d'utiliser l'ensemble des sinistres (Clos et RBNS)	• Hypothèses fortes sur le développement des sinistres non clos
Modèles paramétriques	GLM individuel	• Intégration possible des covariables • Interprétabilité du modèle	• Hypothèse de loi et de forme linéaire • Difficulté à intégrer des covariables qui dépendent du temps

Référence: https://www.mdpi.com/2227-9091/7/3/79/pdf

Implémentation de quelques modèles individuels

Modèles non-paramétriques – utilisation de machine learning (1/2)

- **Principe général**

Ce groupe de modèles permet de prédire l'ultime payé en fonction des covariables, sans avoir à imposer une forme structurelle fixe pour les paiements et sans faire intervenir des états de développement.

Deux problématiques majeures doivent être traitées lors de la construction de ce type de modèles:

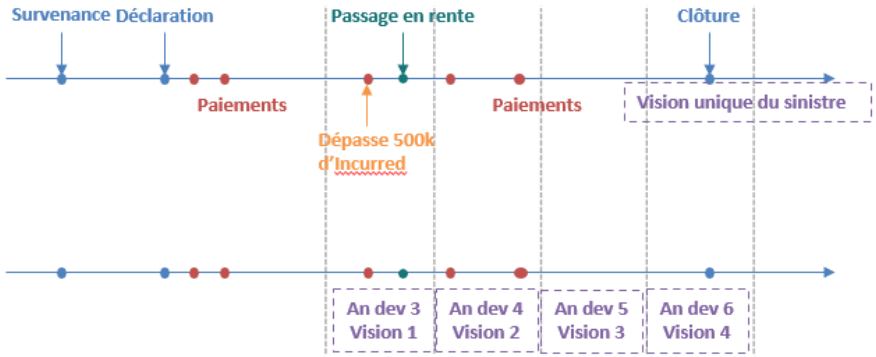
1. Les covariables peuvent n'être que très peu prédictives (c'est le cas sur les données Groupama) → **nécessité de rechercher d'autres covariables et/ou de prendre en compte la volatilité/loi des ultimes par classe**
2. On a besoin d'une fonction réponse à apprendre (ici, l'ultime), qui n'est disponible que pour les sinistres Clos, ce qui induit un biais dans le modèle → **nécessité de compléter la base de données avec des « pseudo-ultimes » pour les sinistres non clos**

- **Possibilité d'enrichissement des covariables** $\{X^{(k)}\}$

Au lieu de considérer un sinistre uniquement à sa dernière vision disponible, on va utiliser toutes les visions du sinistre à chaque année de développement (voir schéma ci-contre).

On dispose alors, pour chaque vision du sinistre, de covariables supplémentaires:

- Dossier passé en rentes ou non
- Temps depuis le passage en rentes
- Nombre de paiements passés
- Montant cumulé payé
- Temps depuis l'ouverture du sinistre
- Temps depuis le dépassement du seuil de 500k
- Provision dossier / dossier



Implémentation de quelques modèles individuels

Modèles non-paramétriques – utilisation de machine learning (2/2)

Modèle A - La solution la plus simple consiste à entraîner le modèle (à définir, de type Random Forest ou XGBoost) sur des données où **seuls les sinistres clos sont inclus**. Ainsi, la réponse est connue pour tous les sinistres.

→Biais de sélection car les sinistres qui sont déjà clos à la date T de provisionnement peuvent avoir des développements plus courts

→Les sinistres dont le développement est plus court peuvent avoir une tendance à avoir des montants totaux payés plus faibles

→Une proportion importante de sinistres est retirée de l'analyse (40%), ce qui entraîne une perte d'informations

Par conséquent, le modèle est presque exclusivement entraîné sur des sinistres avec des réponses d'entraînement plus faibles, ce qui conduit à une sous-estimation des montants totaux payés pour les nouveaux sinistres.

Modèle A: $\widehat{f_A}$ entraîné sur $\left\{\left(X^{(k)}, U^{(k)}\right)\right\}_{\{k \in Clos\}}$ → prédiction de l'ultime pour les RBNS: $\widehat{f_A}\left(\left\{X^{(k)}\right\}_{\{k \in RBNS\}}\right)$

Modèle B - Une deuxième approche consiste à compléter les données par des « pseudo-ultimes » pour les RBNS, avant d'entraîner un algorithme de Machine Learning sur l'ensemble des sinistres (Clos et RBNS).

→Pour ce faire, on développe les sinistres qui sont encore ouverts en utilisant les paramètres d'une approche classique telle que celle de Chain Ladder.

→Hypothèses fortes sur l'ultime des RBNS: on se rajoute une erreur due à l'erreur d'estimation du pseudo-ultime des RBNS par MCL

Modèle B: $\widehat{f_B}$ entraîné sur $\left\{\left(X^{(k)}, U^{(k)}\right)\right\}_{\{k \in Clos\ \&RBNS\}}$, où

- $\left\{U^{(k)}\right\}_{\{k \in RBNS\}}$ est obtenu via le développement de triangle par MCL
- $\left\{U^{(k)}\right\}_{\{k \in Clos\}}$ est connu

Remarque: sont implémenté, pour chacun des deux modèles A et B, deux version, **modèle A (resp. B) global**, et **modèle A (resp. B) différencié** (2 modèles sont construits, un pour les visions de sinistres inférieures à 500k, et un pour les visions de sinistres supérieures à 500k)

Par construction, on surestime l'ultime avec les modèles « différenciés » car on applique le « mauvais « modèle » aux premières années de vie du sinistre.

Implémentation de quelques modèles individuels

Modèles paramétriques – processus de Poisson

- **Principe général**

Utilisation d'un modèle individuel GLM quasi-Poisson pour estimer l'Ultime. Une function de lien logarithme est utilisée et les coefficients sont estimés en maximisant la function de vraisemblance de Poisson. Le modèle est calibré sur l'ensemble des sinistres disponibles, en considérant les covariables et les paiements incrémentaux.

- **Détail de formalisation / spécification du modèle individuel – Modèle C**
  - On suppose que les paiements incrémentaux $Y_j^{(k)}$ (paiement de l'année de développement j du sinistre k) suivent une loi GLM caractérisée par le délai de développment et les covariables du sinistre.
  - La prédiction $\widehat{Y_j^{(k)}}$ est obtenue par $\widehat{Y_j^{(k)}} = g^{-1}\left(x_j^{(k)}\hat{\beta}\right)$, avec $\hat{\beta}$ l'estimateur du maximum de vraisemblance de β (le vecteur des paramètres du modèle).
  - On peut ensuite évaluer, pour chaque sinistre du portefeuille et chaque période de développement, une réserve individuelle, et on en déduit ainsi la réserve totale pour les RBNS.
  - Quelques résultats d'implémentation sur les données Groupama:
    - Différentes variantes ont été testées (différentes lois, utilisation de stepAIC et stepBIC, modélisation des charges incrémentales et des paiements incrémentaux
    - L'analyse des résidus ne permet pas de valider les modèles.

Implémentation de quelques modèles individuels

Résultats obtenus et interprétations

	Erreurs de prédiction entre l'ultime réel et les différentes prédictions					
Année N de backtesting	Gestionnaire	A - global	A - différencié	B - global	B_ différencié	C
2010	25%	-20%	15%	14%	16%	23%
2011	34%	-10%	21%	10%	26%	33%
2012	41%	-9%	23%	12%	29%	50%
2013	38%	-7%	23%	4%	27%	63%
2014	44%	-6%	22%	8%	21%	74%
2015	29%	-17%	14%	7%	21%	78%
2016	32%	-14%	10%	31%	22%	92%
2017	28%	-4%	15%	30%	29%	112%
2018	21%	-8%	14%	31%	31%	510%

Estimations* (Mds€)	Gestionnaires	A global	A différencié	B global	B différencié	C	Chain Ladder	Chain Ladder alternatif
Ultimes	2,9	1,6	1,8	2,2	2,4	4,9	2,6	2,4
Provisions	2,2	0,9	1,1	1,5	1,7	4,3	1,9	1,8

Hypothèse: les gestionnaires sinistres surestiment l'ultime d'au moins 20%
→ l'ordre de grandeur obtenu en retraitant la provision des gestionnaires de 20%, coïncide avec celui obtenu avec le modèle B différencié et le modèle CL retraité.

*** Note:** les chiffres présentés ici ont été modifiés par souci de confidentialité

Enjeux et limites du passage à un modèle individuel (1/2)

Aspects normatifs

- Capacité de calibrer une volatilité plus précise et donc une marge de risque plus pertinente sur la sinistralité grave
- Normes de la sinistralité grave et projections du plan stratégique (volet auto RC graves) à actualiser
- Possibilité d'enrichir le modèle de provisionnement avec des données complémentaires et/ou avec des règles de gestion de sinistres
- Nécessité de garantir la stabilité des méthodologies dans le temps

Enjeux et limites du passage à un modèle individuel (2/2)

Aspects opérationnels

- Mise en place d'un suivi rapproché de la sinistralité
- Fluidité des échanges entre les différents métiers autour du provisionnement
- Meilleure gestion des marges et/ou déficits potentiels de provisionnement
- Nécessité de développer plusieurs modèles (RBNS, IBNR, IBNR XS) pour le provisionnement des graves
- Maintenance des modèles dans le temps.

Perspectives et conclusions

- La simple exploitation des données individuelles, sans forcément passer à un modèle de provisionnement individuel, fournit des informations exploitables et utiles pour l'activité
- Les provisions à effectuer ne sont pas les mêmes selon les caractéristiques des sinistres → on ne peut pas le voir forcément dans les méthodes agrégées
  - Les modèles RF donnent des résultats similaires aux méthodes agrégées
  - A ce stade, pas de modèle parfait → chercher d'autres covariables etc
- Le provisionnement individuel nécessite un travail important sur la qualité des données
- Il n'y a pas un seul modèle de provisionnement individuel, il faut construire celui s'adapte le plus aux données individuelles disponibles