OPENCLASSROOMS

P5 : Segmentez des clients d'un site e-commerce

22/06/2022

DUBART Maxime

Segmentation clients e-commerce

OBJETIFS

Identifier les différents types d’utilisateurs

Bons vs. moins bons clients en termes de commandes et satisfaction

Analyse de stabilité des segments

Fréquence de mise à jour et contrat de maintenance

Segmentation clients e-commerce

METHODE

RFM (récence, fréquence, montant)

Récence : temps depuis le dernier achat
Fréquence : fréquence des achats sur la période
Montant : montant des achats sur la période

Extension (RFMe)

Satisfaction : satisfaction moyenne

Segmentation clients e-commerce

DONNEES

Données disponibles

Récence | Fréquence | Montant | Satisfaction

Orders : table des commandes

Champs d'intérêt : *order_id, customer_id, order_status, order_approved_at*

Customers : table des clients

Champs d'intérêt : *customer_id, customer_unique_id*

Payments : table des paiements (possiblement fractionnés)

Champs d'intérêt : *order_id, payment_value*

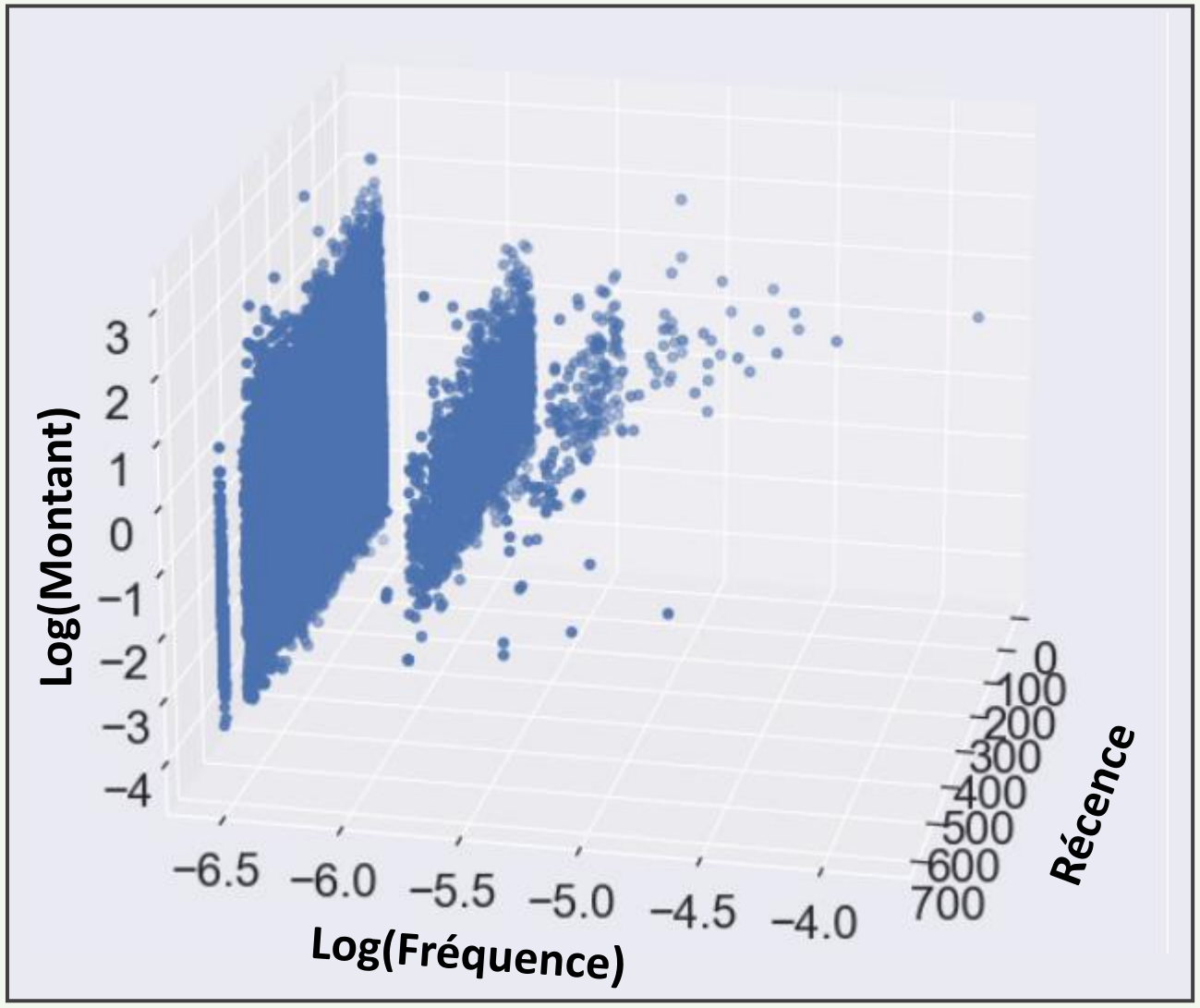
Reviews : table des avis / notes clients par commande

Champs d'intérêt : *order_id, review_score*

Segmentation clients e-commerce

DONNEES

RFM

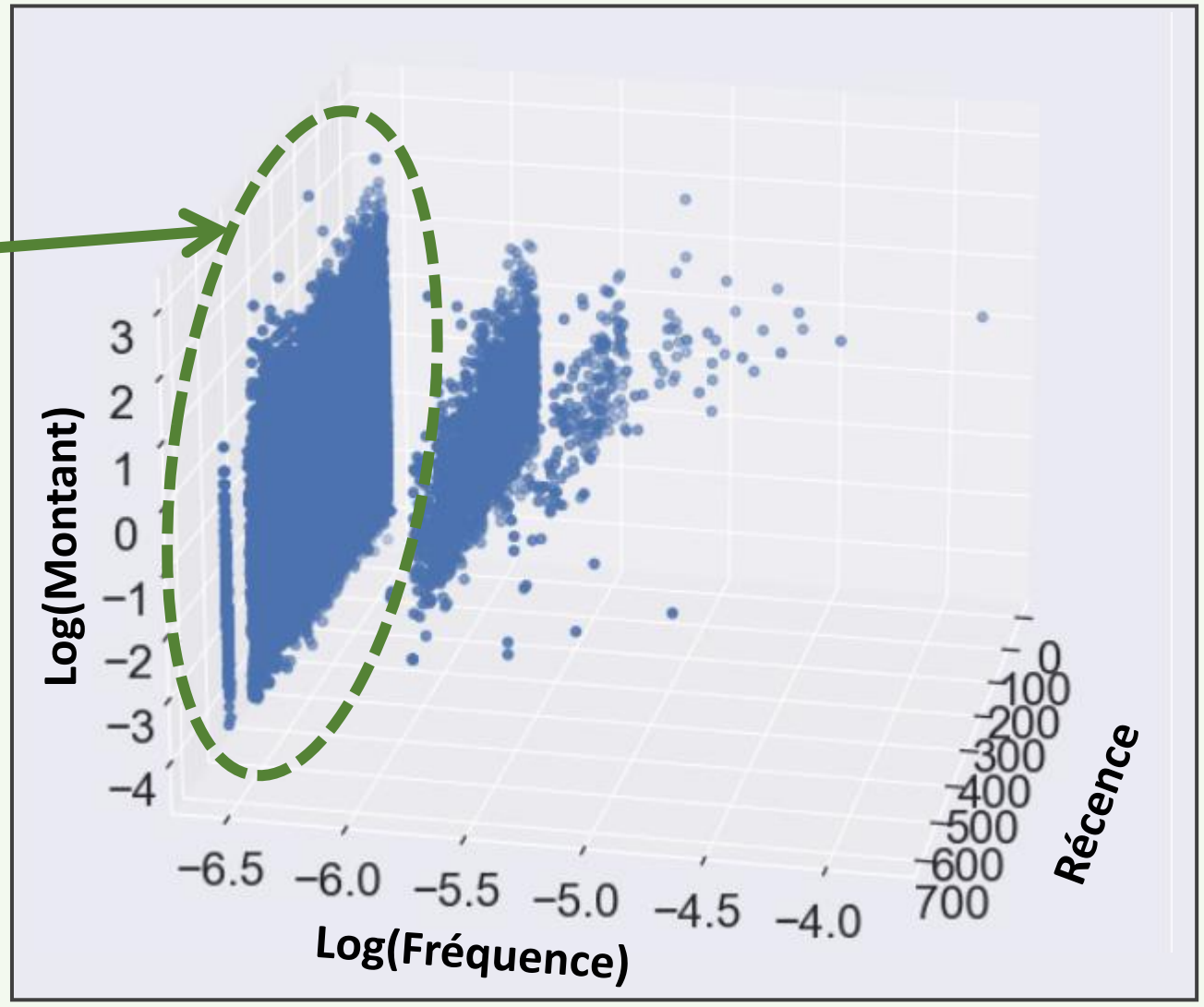


Segmentation clients e-commerce

DONNEES

RFM

Clients avec une commande



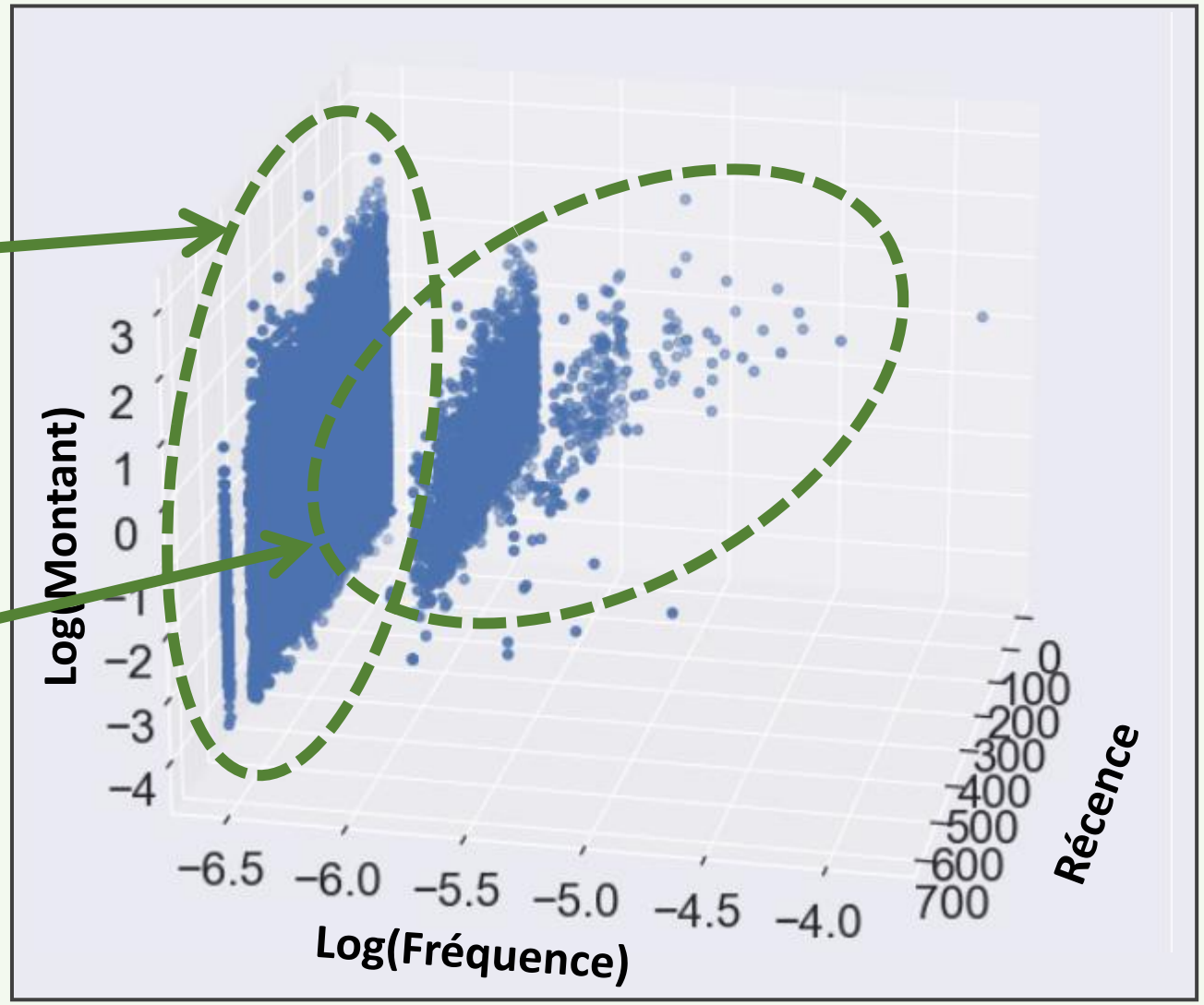
Segmentation clients e-commerce

DONNEES

RFM

Clients avec une commande

Clients avec plus d'une commande (env. 3%)



Segmentation clients e-commerce

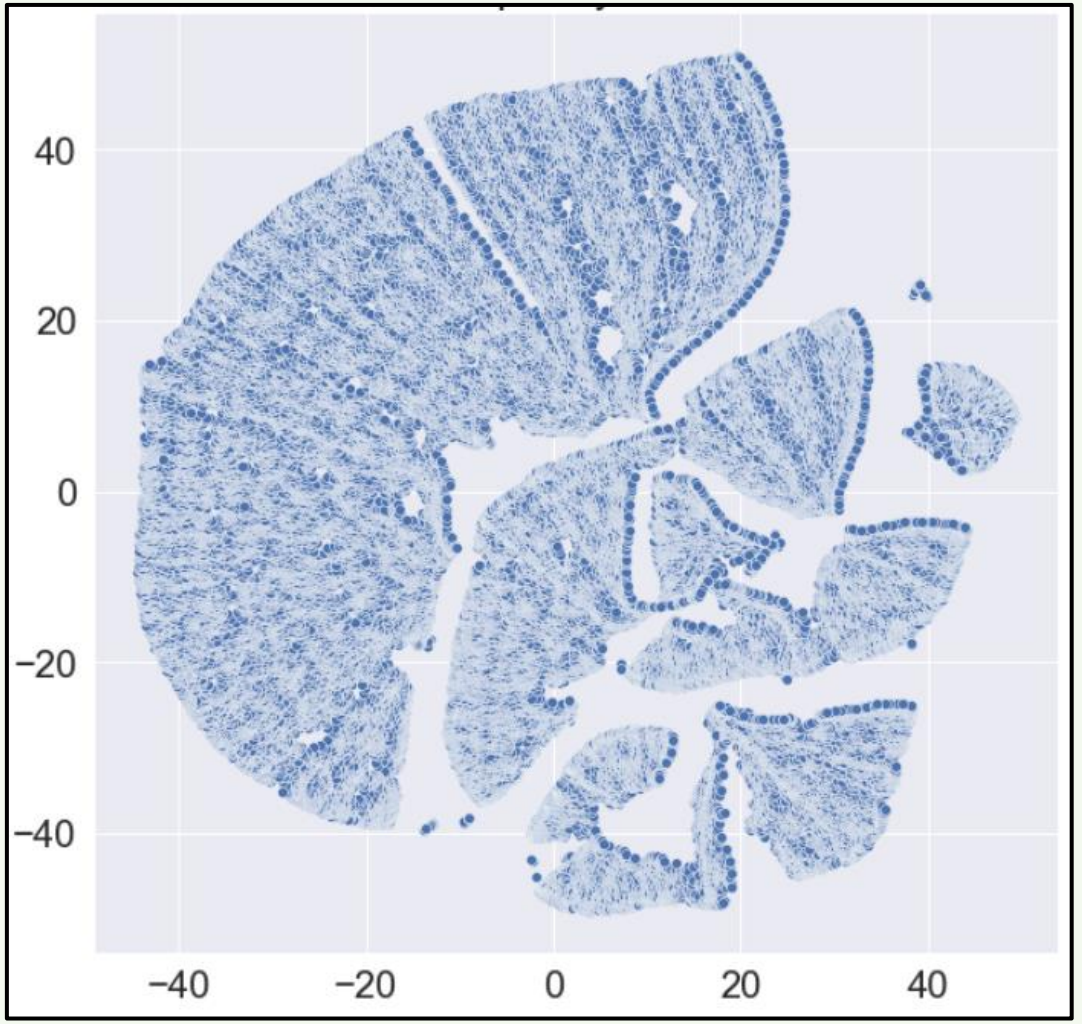
DONNEES

RFMe

Difficile à représenter en 4D
Réduction de dimension (t-SNE)

Pas de segmentation claire

Perplexity : 200



Segmentation clients e-commerce

CLUSTERING - RFMe

Méthodes

K-means

Clustering hiérarchique

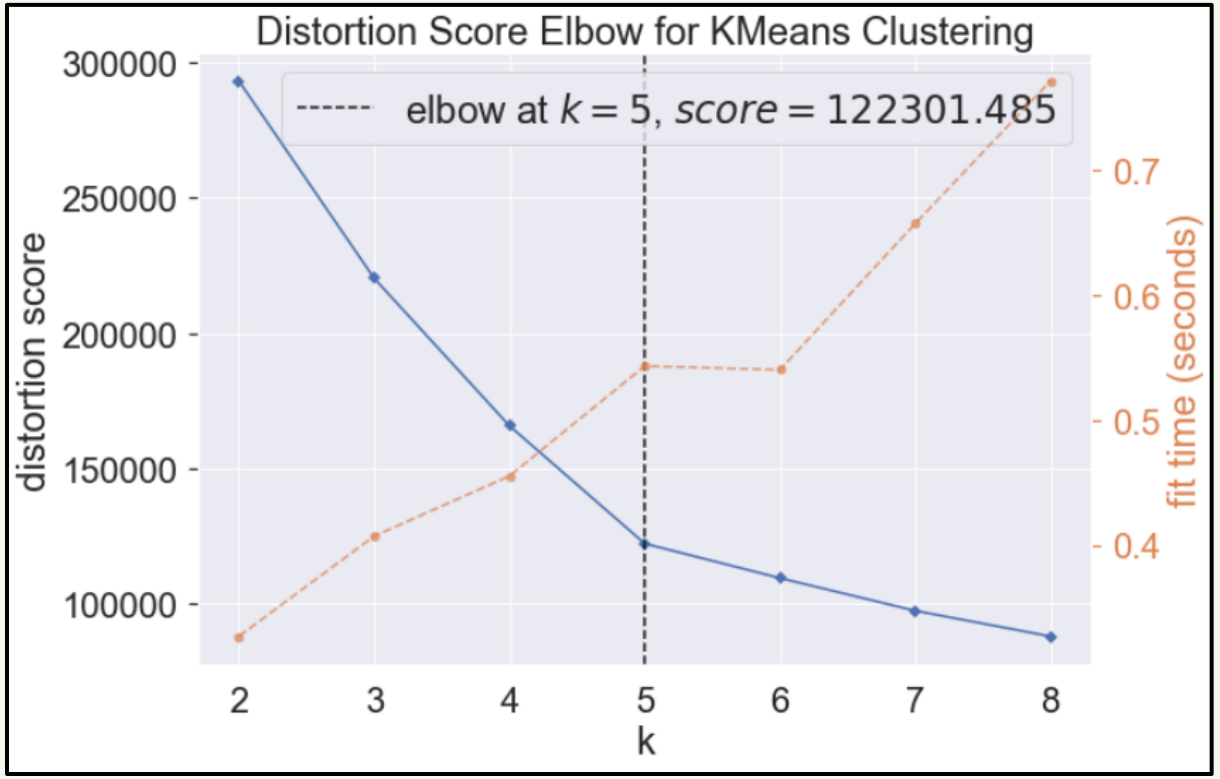
DBScan

Segmentation clients e-commerce

k-means

Meilleur **k** *a priori* : k = 5

Distortion (wcss) : $\sum_{i}^{N} d(x_i, K(i))^2$



Segmentation clients e-commerce

CLUSTERING - RFMe

k-means

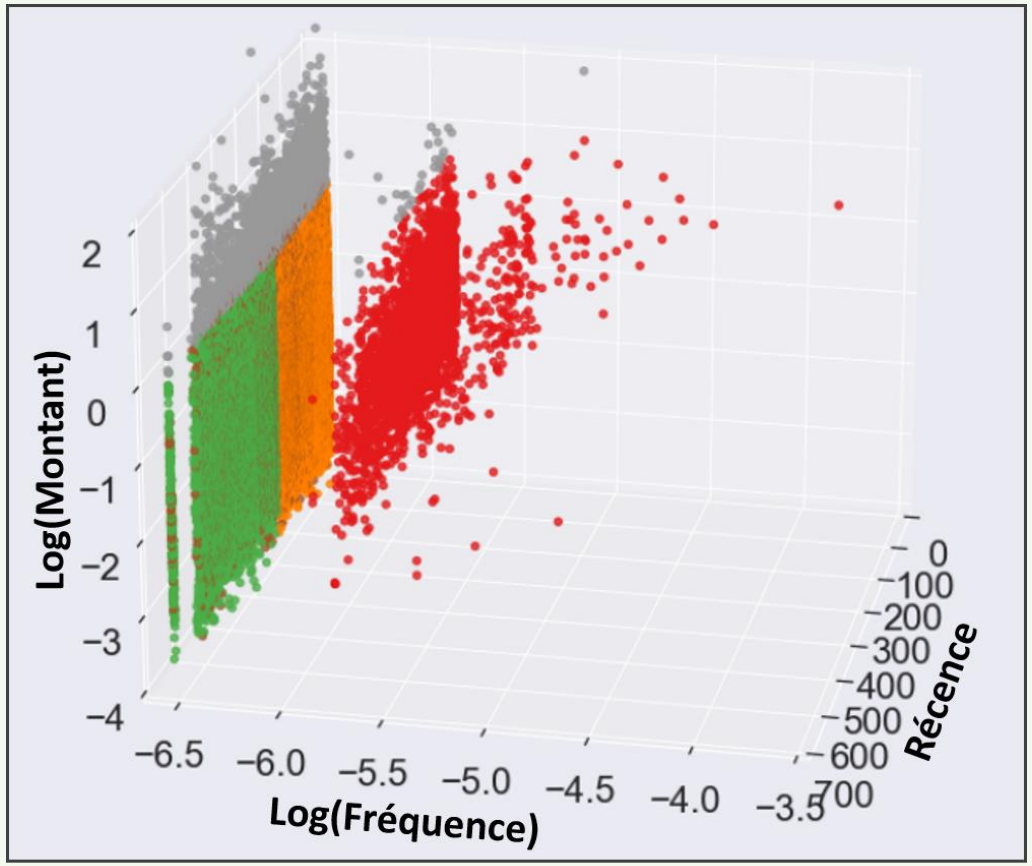
Meilleur **k** *a priori* : k = 5

Une commande

- Groupe #1 (33%): vielle
- Groupe #2 (44%): récente
- Groupe #3 (17%): ?
- Groupe #4 (2%): un montant très important

> une commande

- Groupe #0 (3%)



Segmentation clients e-commerce

CLUSTERING - RFMe

k-means

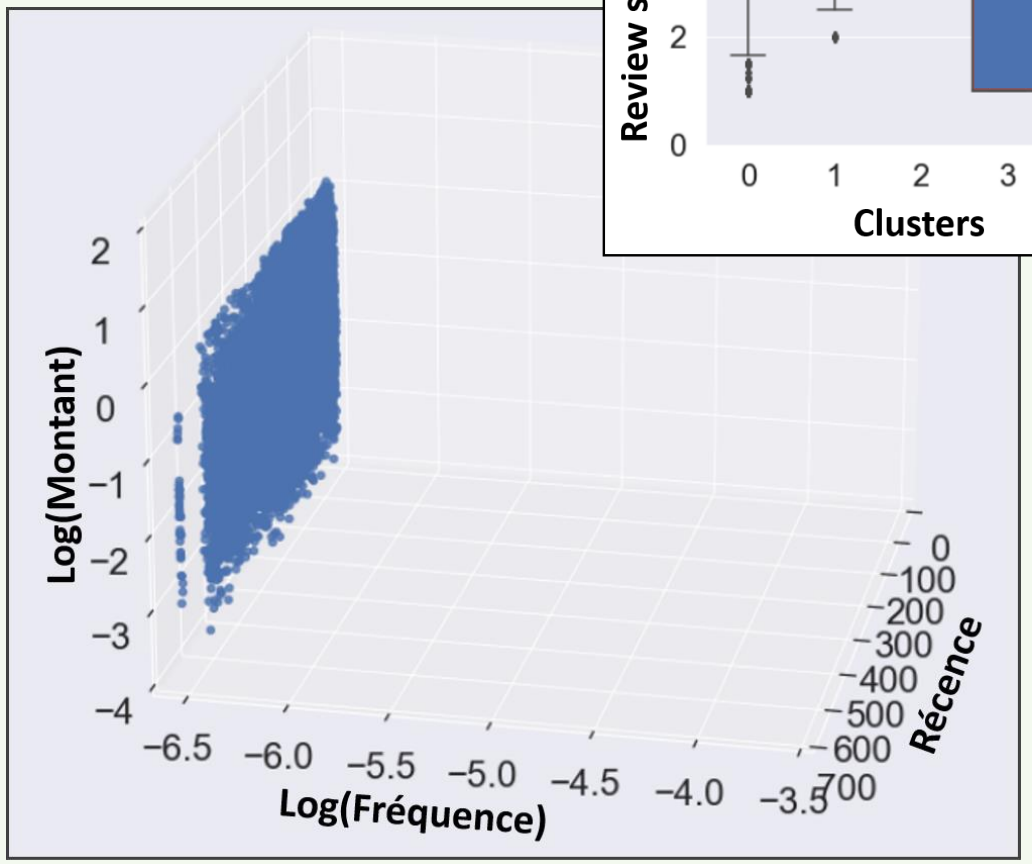
Meilleur **k** *a priori* : k = 5

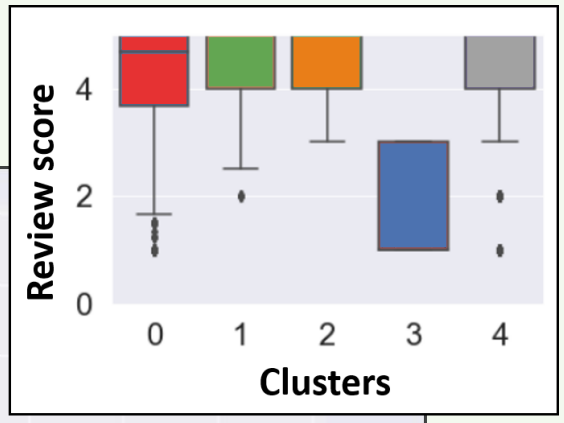
Une commande

- Groupe #1 (33%): vielle
- Groupe #2 (44%): récente
- Groupe #3 (17%): ?
- Groupe #4 (2%): un montant très important

> une commande

- Groupe #0 (3%)





Segmentation clients e-commerce

CLUSTERING - RFMe

k-means

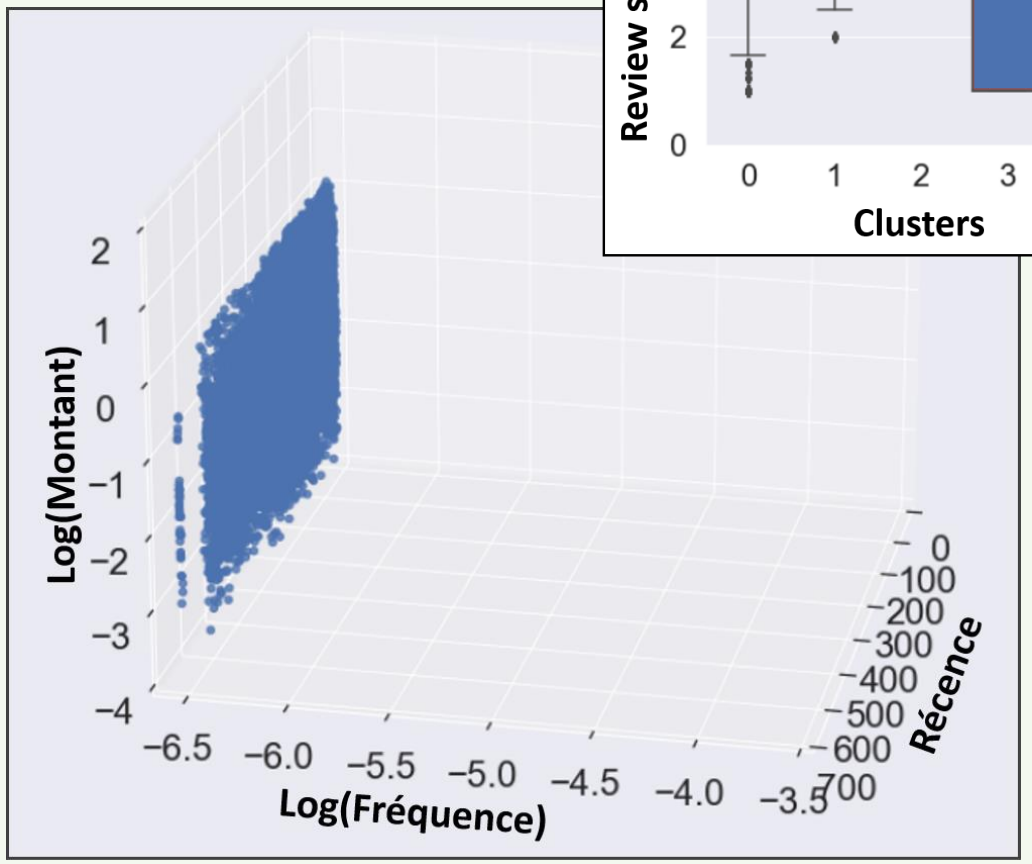
Meilleur **k** *a priori* : k = 5

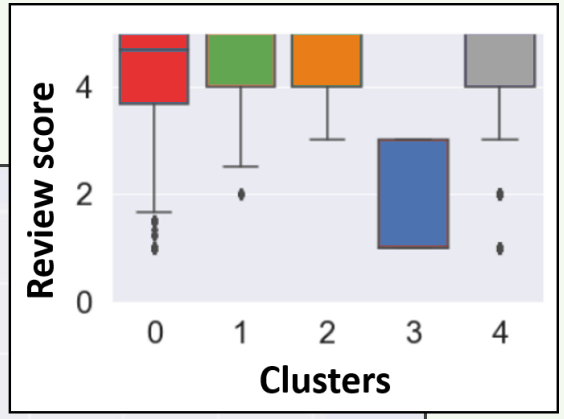
Une commande

- Groupe #1 (33%): vielle
- Groupe #2 (44%): récente
- Groupe #3 (17%): ?
- Groupe #4 (2%): un montant très important

> **une** commande

- Groupe #0 (3%)





Faut-il conserver le groupe #4 ? À voir avec le client

Segmentation clients e-commerce

CLUSTERING - RFMe

k-means

Meilleur **k** *a priori* : k = 5

Une commande
- Groupe #1 (33%): vielle
- Groupe #2 (44%): récente
- Groupe #3 (17%): ?
- Groupe #4 (2%): un montant très important

> **une** commande
- Groupe #0 (3%)

Stabilité

Calcule de l'ARI entre prédictions de modèles entrainés sur des échantillons aléatoires (20% et 30% des données), répété 20x

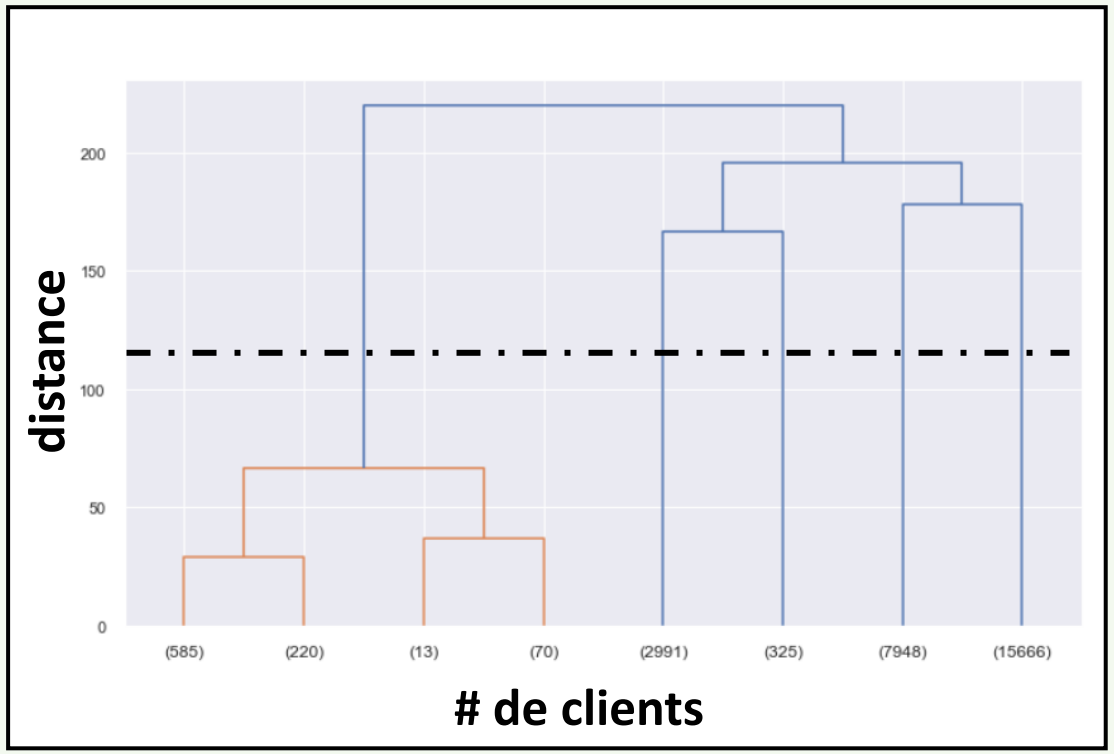
ARI = 0.98 ± 0.0072

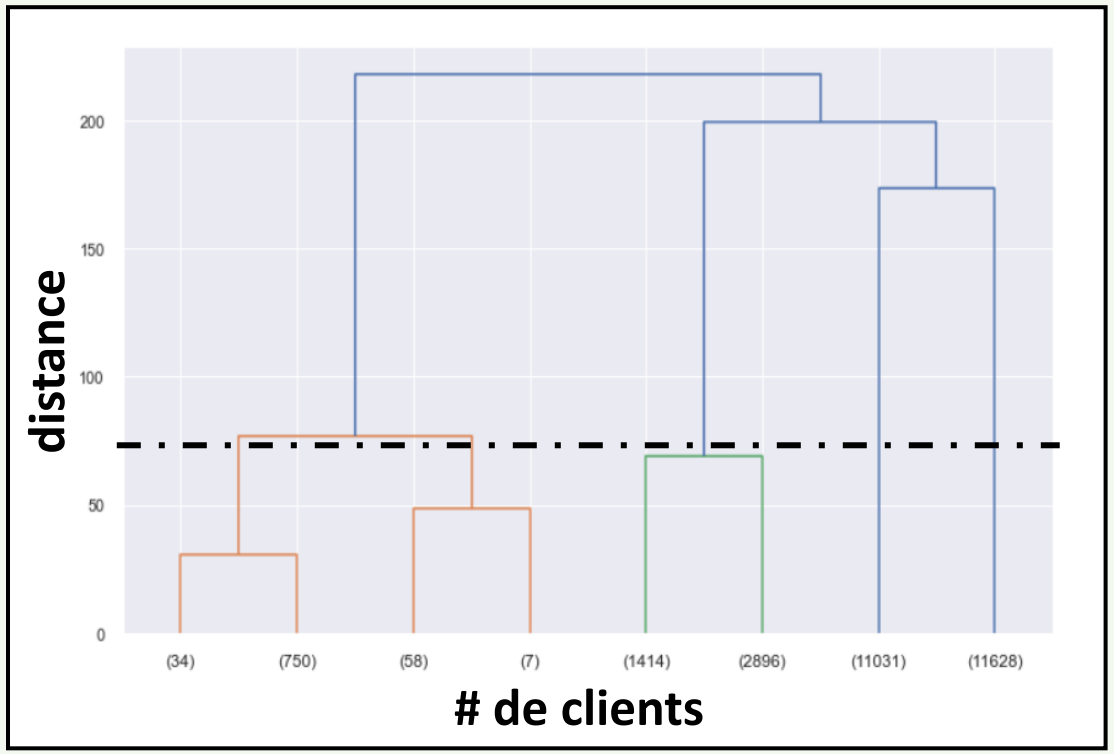
Segmentation clients e-commerce

CLUSTERING - RFMe

Clustering hiérarchique (k=5) – entrainement 30%

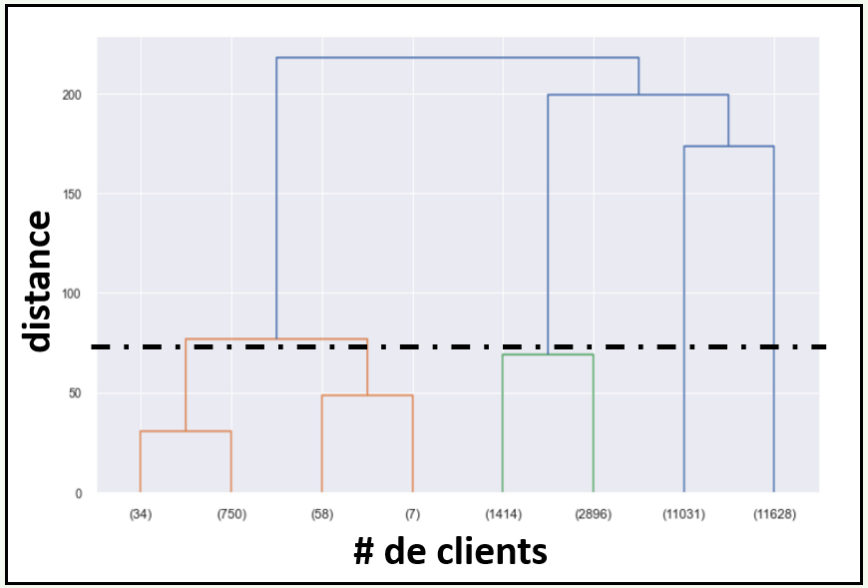
Configurations alternatives





Segmentation clients e-commerce

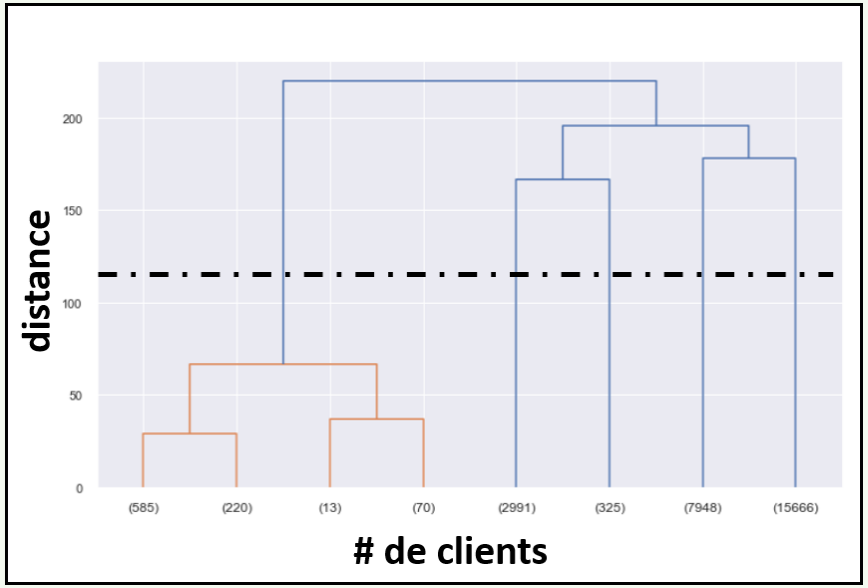
Clustering hiérarchique (k=5)



Stabilité

Calcule de l'ARI entre prédictions de modèles (via knn) entrainés sur des échantillons aléatoires (30% des données), répété 5x

ARI = 0.46 ± 0.15



Comparé au k-means

ARI = 0.56 ± 0.09

Segmentation clients e-commerce

CLUSTERING - RFMe

Clustering hiérarchique (k=5)

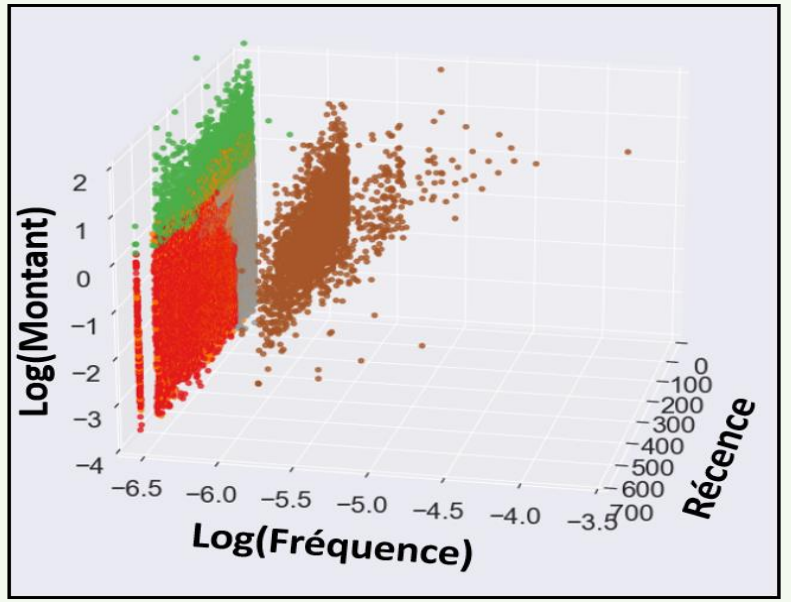
Stabilité

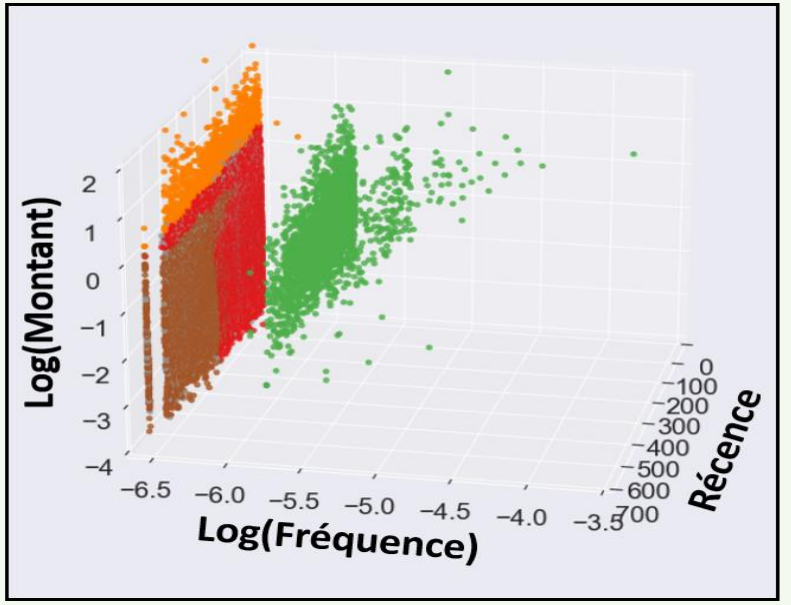
Calcule de l'ARI entre prédictions de modèles (via knn) entrainés sur des échantillons aléatoires (30% des données), répété 5x

ARI = 0.46 ± 0.15

Comparé au k-means

ARI = 0.56 ± 0.09



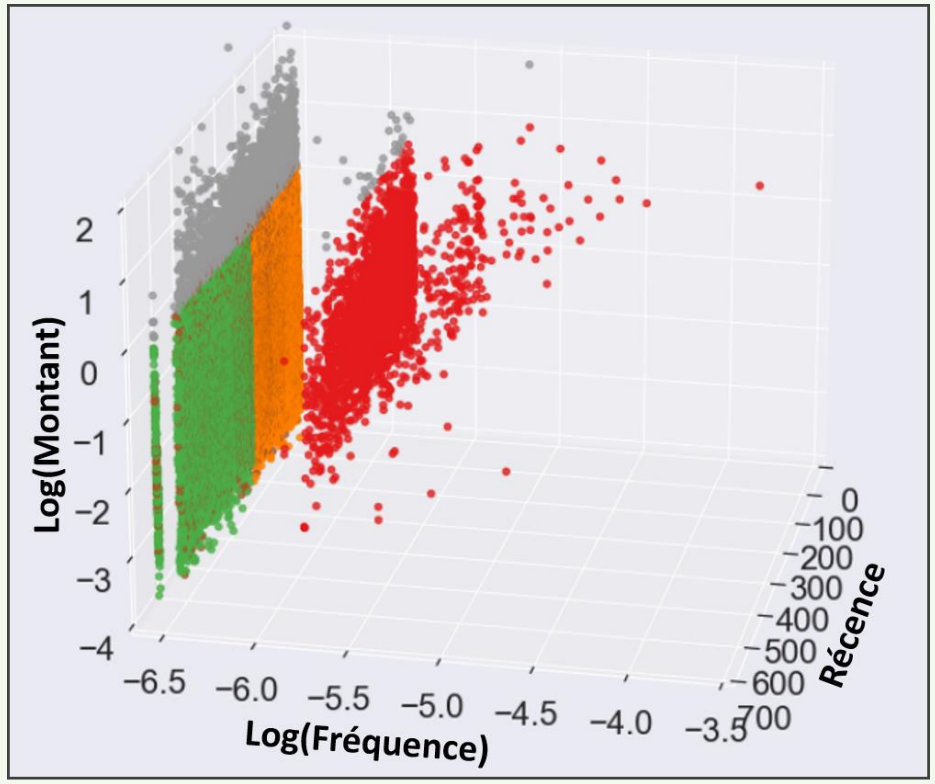


Segmentation clients e-commerce

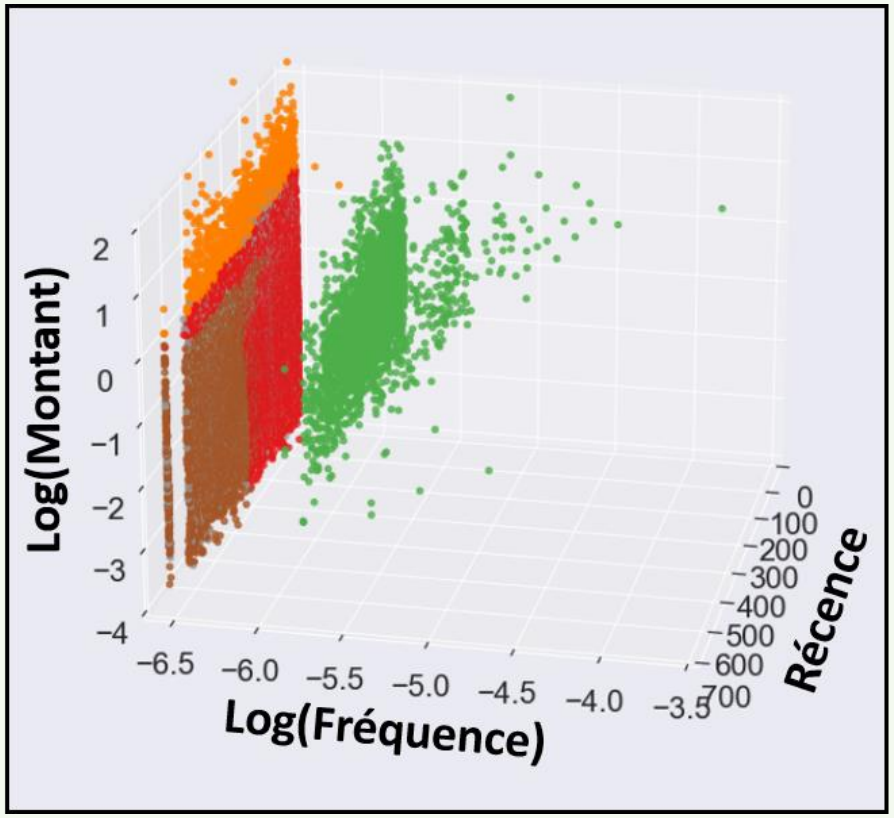
CLUSTERING - RFMe

Clustering hiérarchique (k=5)

K-means



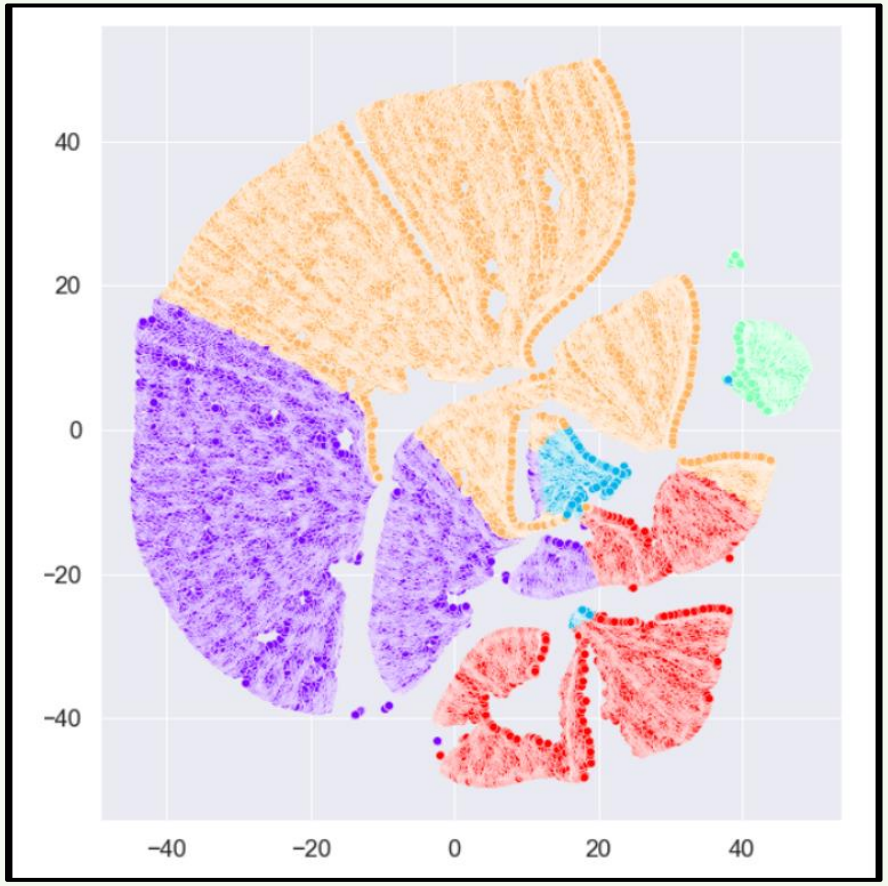
Hierarchical clustering



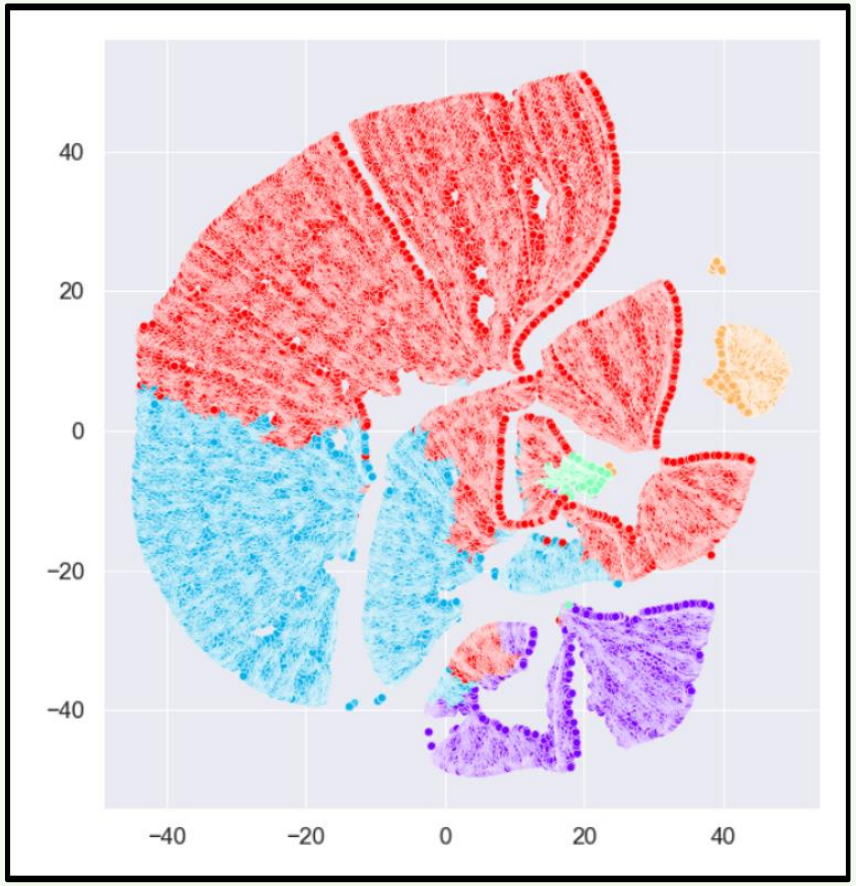
Segmentation clients e-commerce

Clustering hiérarchique (k=5)

K-means



Hierarchical clustering



Segmentation clients e-commerce

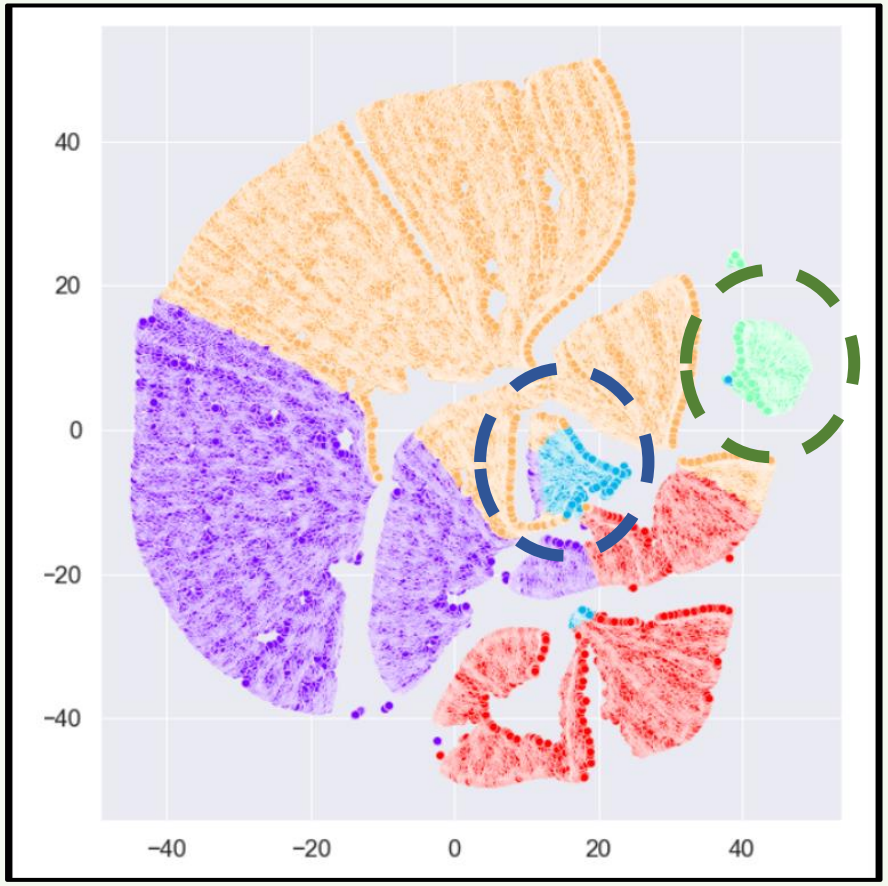
CLUSTERING - RFMe

Clustering hiérarchique (k=5)

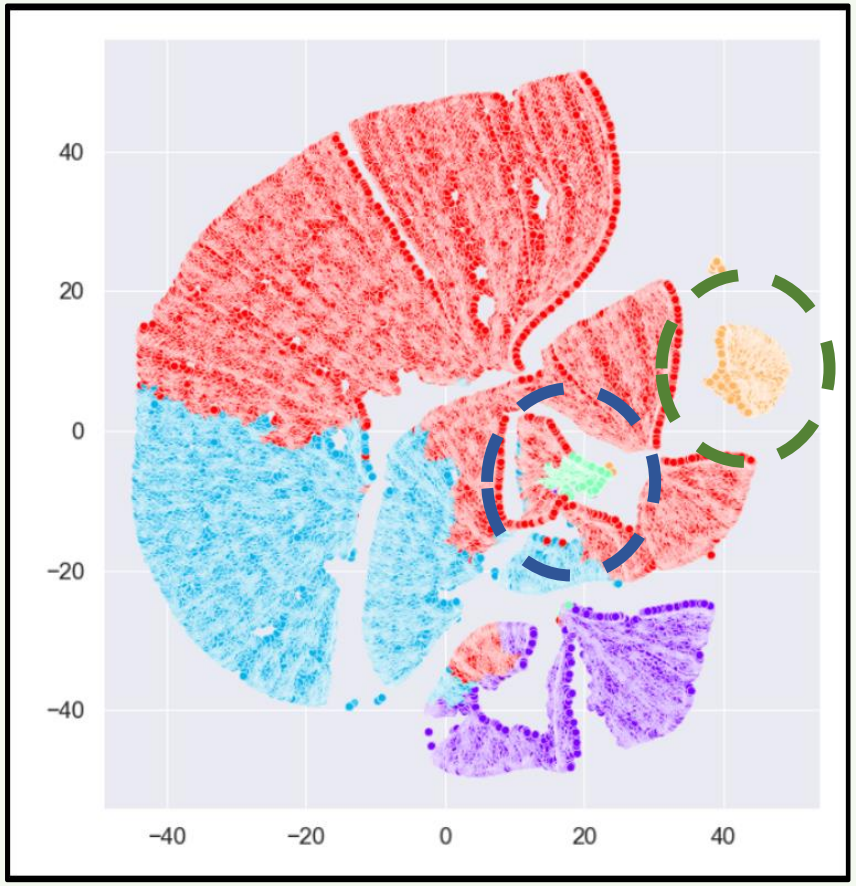
Bons clients

Montants importants

K-means



Hierarchical clustering



Clustering hiérarchique (k=5)

Segmentation moins claire / moins stable que k-means

Ne permet pas (par défaut) la segmentation de nouveau clients

Difficile à utiliser sur grand jeu de données

DBscan

Heuristique pour déterminer les hyper-paramètres (ε, minPts)

ε : calculer pour chaque point de l'espace la distance à son plus proche voisin. Prendre **ε** tel qu'une part « suffisamment grande » des points aient une distance à son plus proche voisin inférieure à **ε**

minPts : calculer pour chaque point le nombre de ses voisins dans un rayon de taille **ε** (la taille de son **ε** -voisinage). Prendre MinPts tel qu'une part « suffisamment grande » des points aient plus de MinPts points dans leur **ε** -voisinage.

(**ε, minPts**) = (0.046, 1)

754 clusters identifiés … Dbscan non adapté (données non homogènes, pas de distinction claire)

Conlusion sur méthode clustering

K-means : produits segments stables, répétables, logiques

Hclust : segmentation instable

DBScan : inadapté

Segmentation clients e-commerce

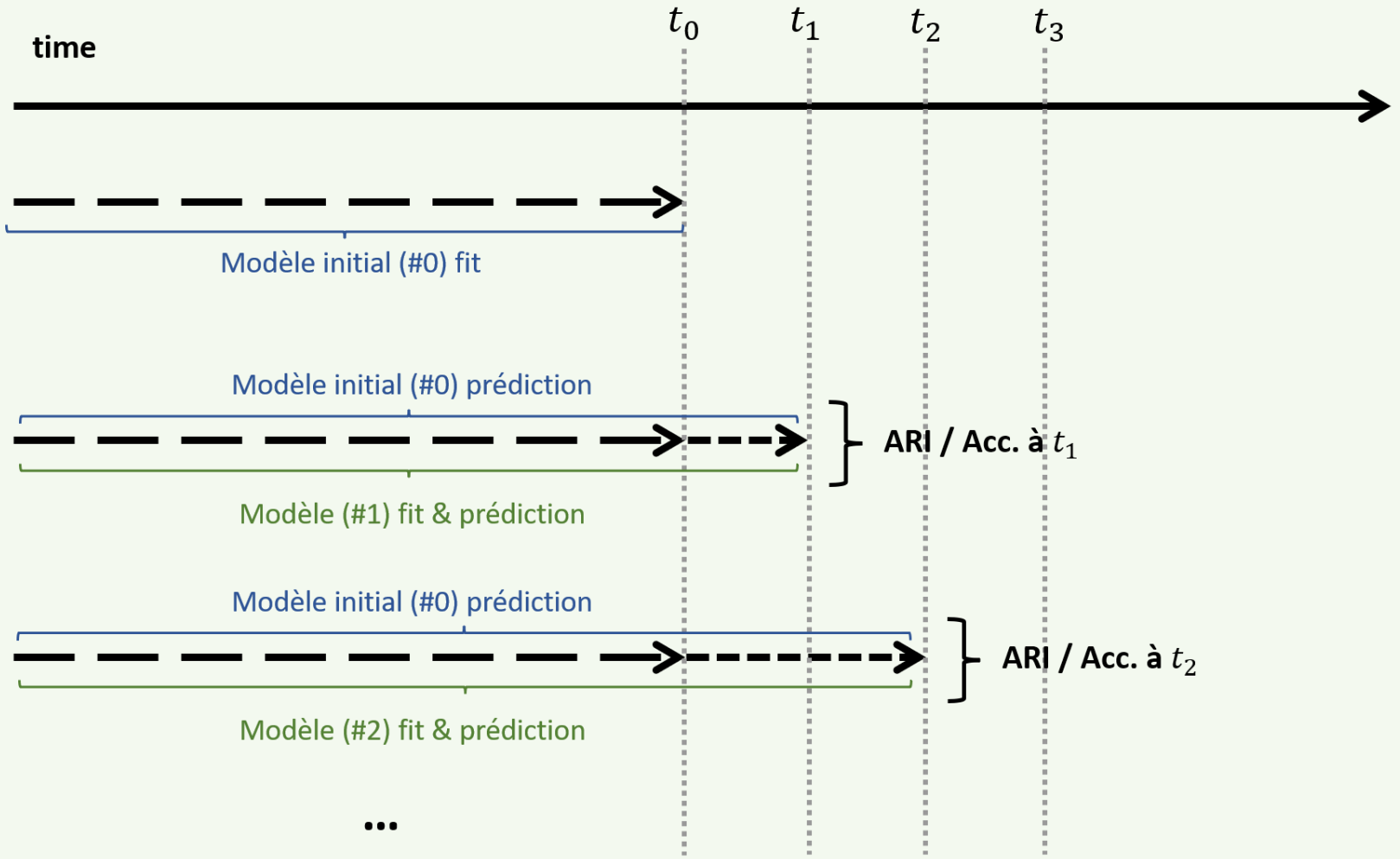
Simulartion maintenance

Fréquence de mise à jour ?

Segmentation clients e-commerce

Simulartion maintenance

Fréquence de mise à jour ?

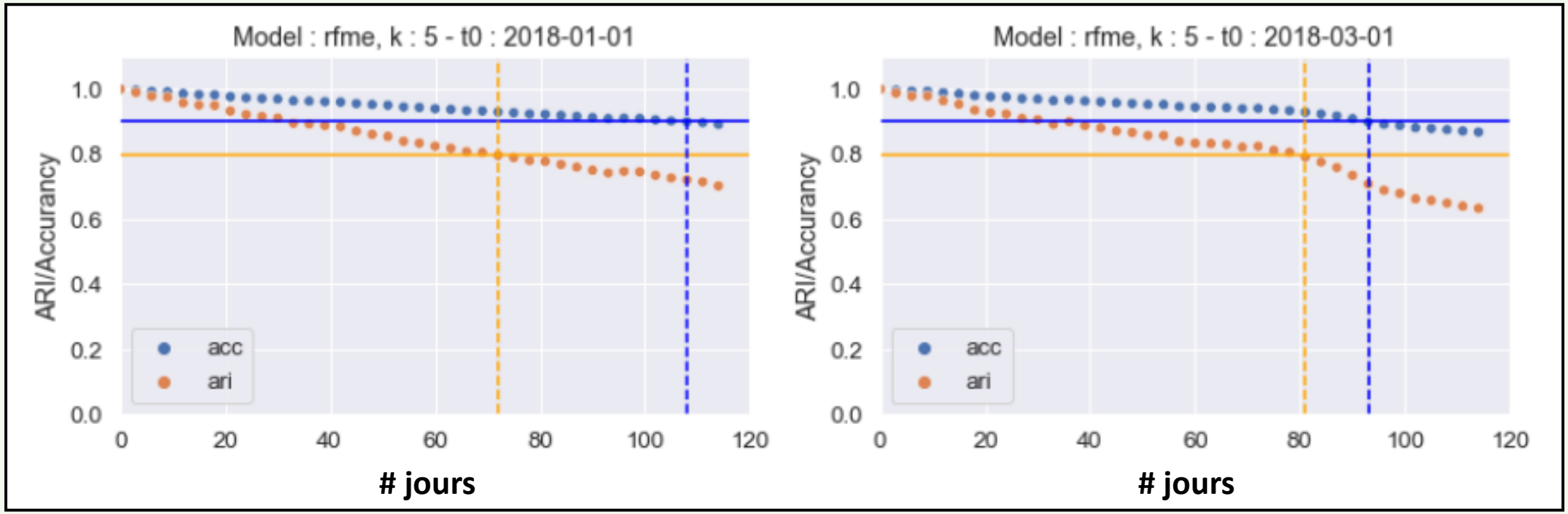


A quel t_x l'ARI < 0.8 (acc. < 0.9) ?

Segmentation clients e-commerce

Simulartion maintenance

Fréquence de mise à jour ?



A quel t_x l'ARI < 0.8 (acc. < 0.9) ?

Avec ARI : environ 70/80 jours.

Avec Accuracy : environ 100 jours.

Segmentation clients e-commerce

Segments instables ?

Segment « bons clients » : très stable

Autres segments : transferts en partie artéfactuels (liés à la standardisation)

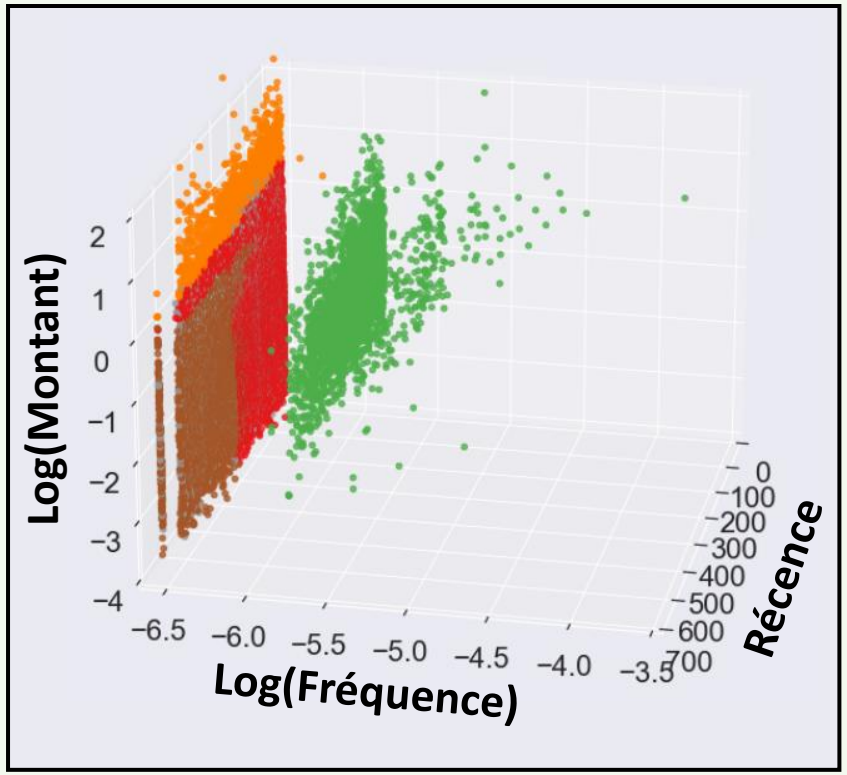
Segmentation clients e-commerce

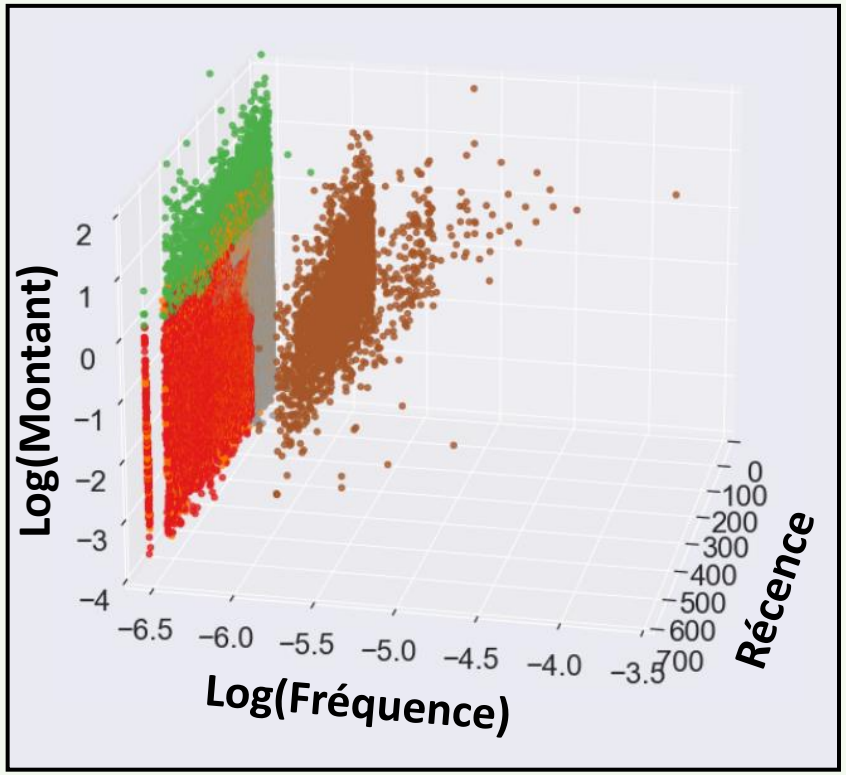
Conclusion

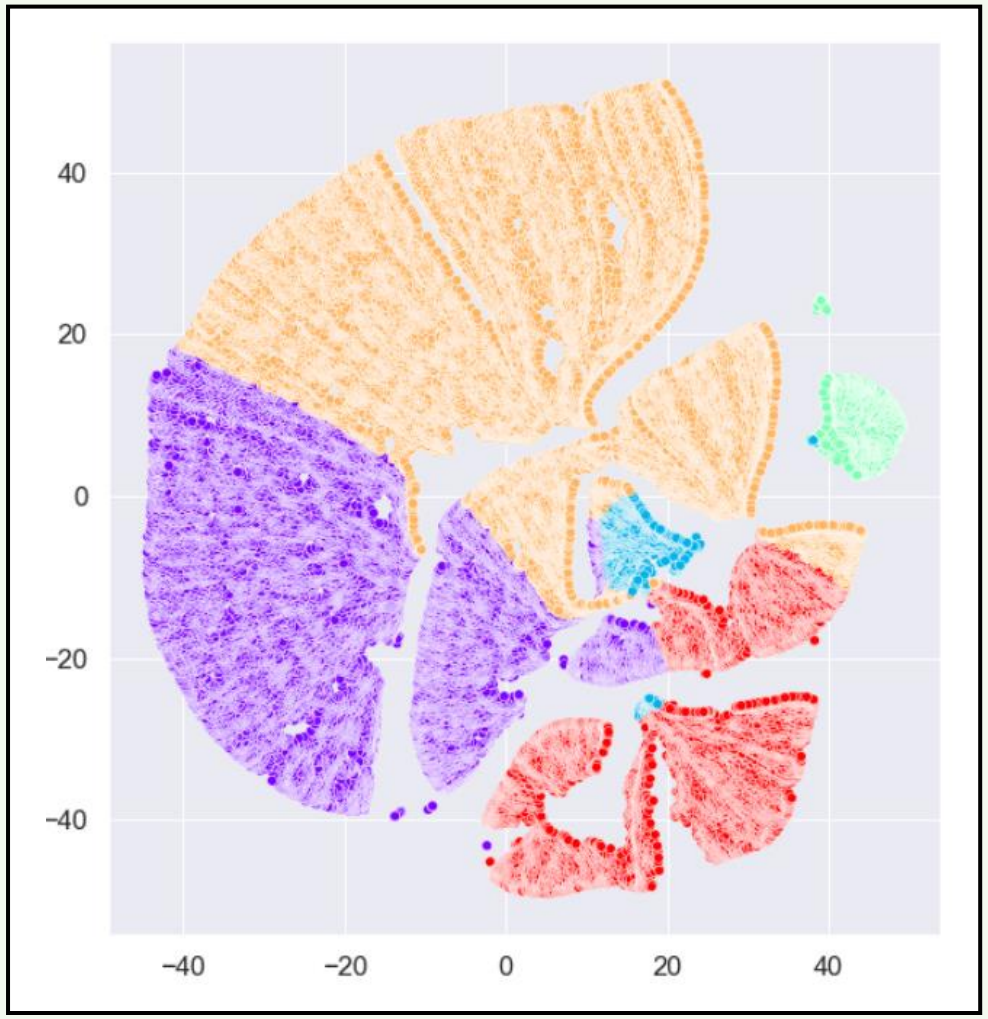
Classification **k-means** permet de segmenter le clients en **5 catégories**, facile à décrire (clients avec commande récente / ancienne, avec montant important, avec une fréquence importante, ou mécontents).

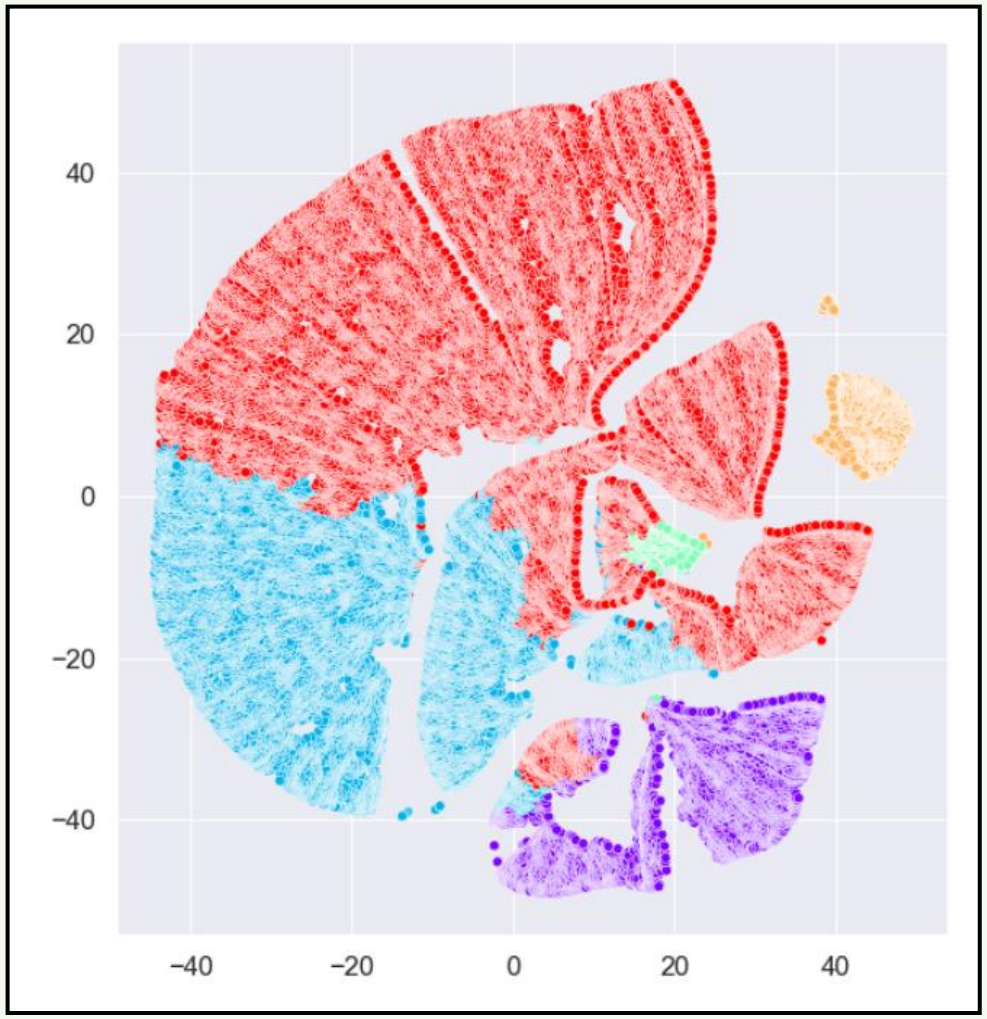
Une **maintenance** tous les 2/3 mois permet d'assurer la stabilité des segments.

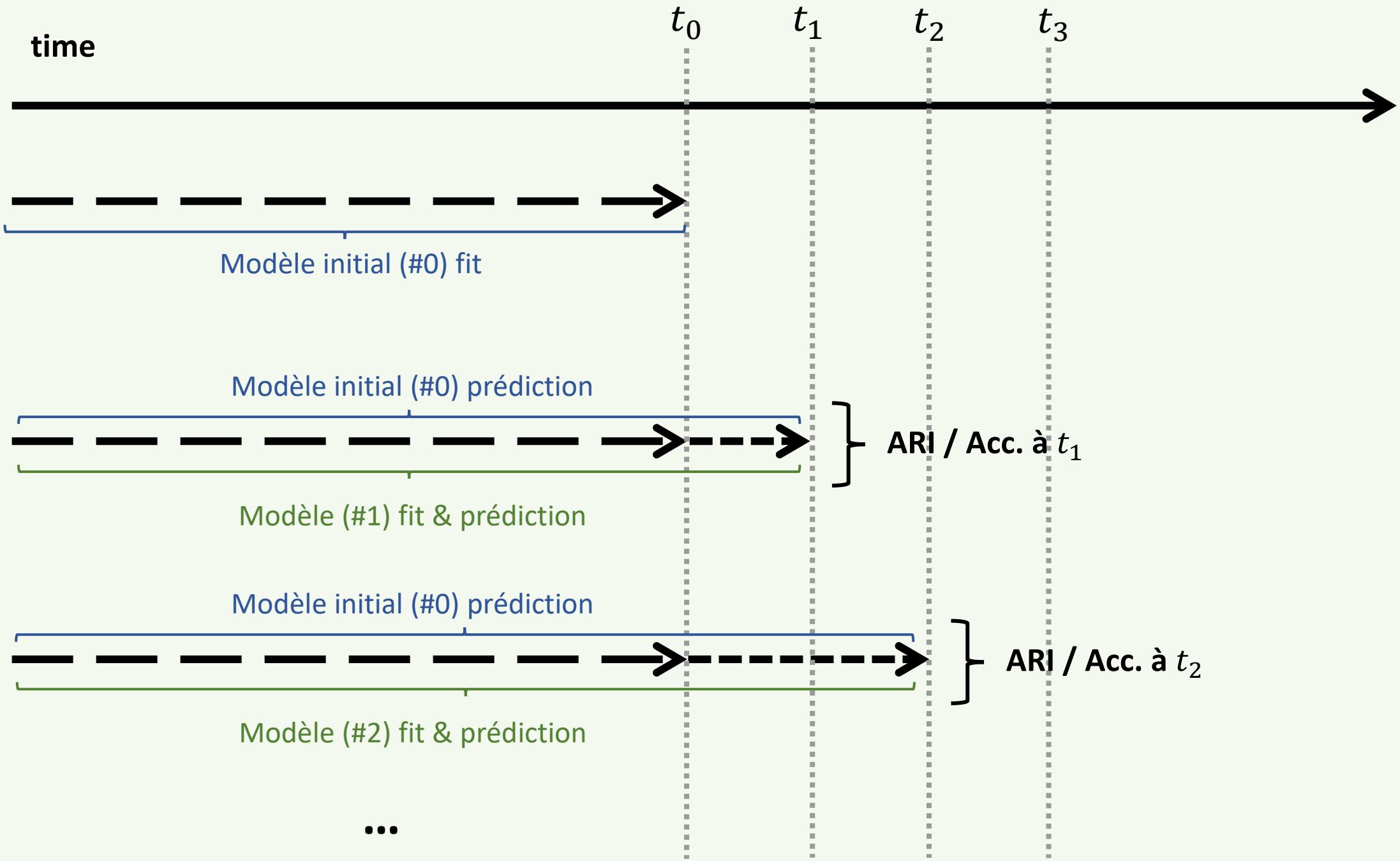
Merci

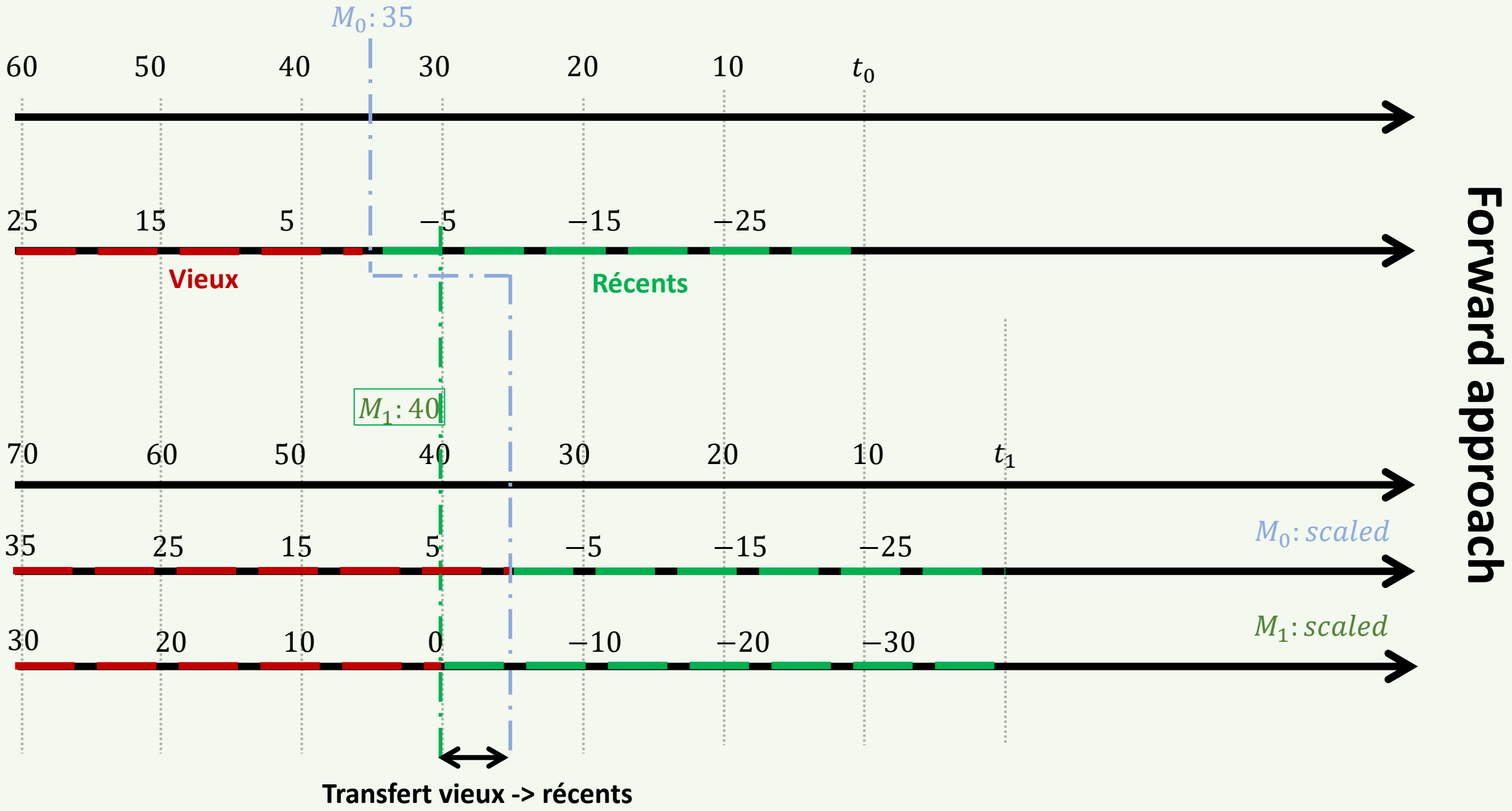

Log(Montant)
2
1
0
−1
−2
−3
−4
−6.5
−6.0
−5.5
−5.0
−4.5
−4.0
−3.5
Log(Fréquence)
0
−100
−200
−300
−400
−500
−600
−700
Récence

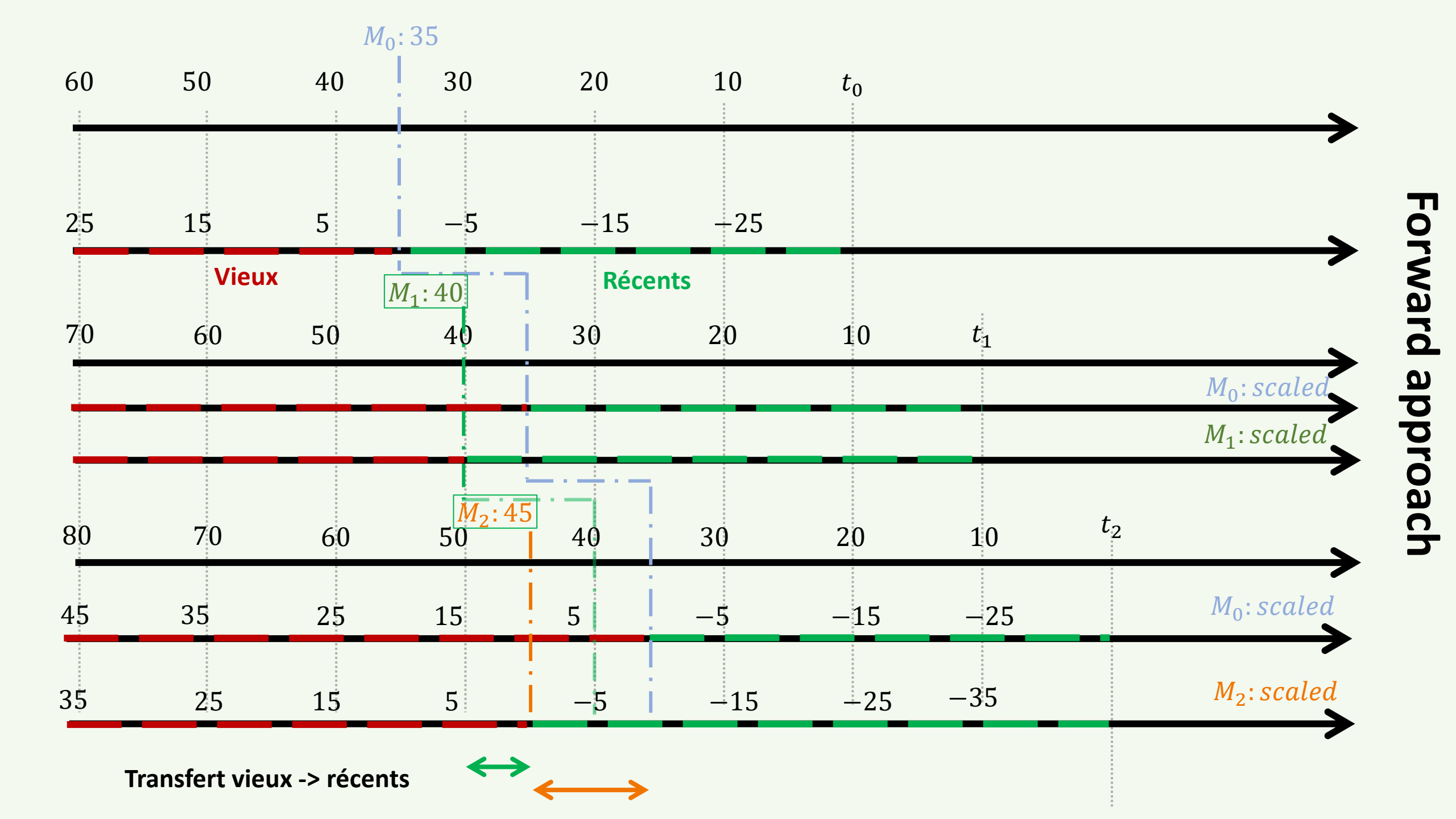

Log(Montant)
2
1
0
−1
−2
−3
−4
−6.5
−6.0
−5.5
−5.0
−4.5
−4.0
−3.5
Log(Fréquence)
0
−100
−200
−300
−400
−500
−600
−700
Récence

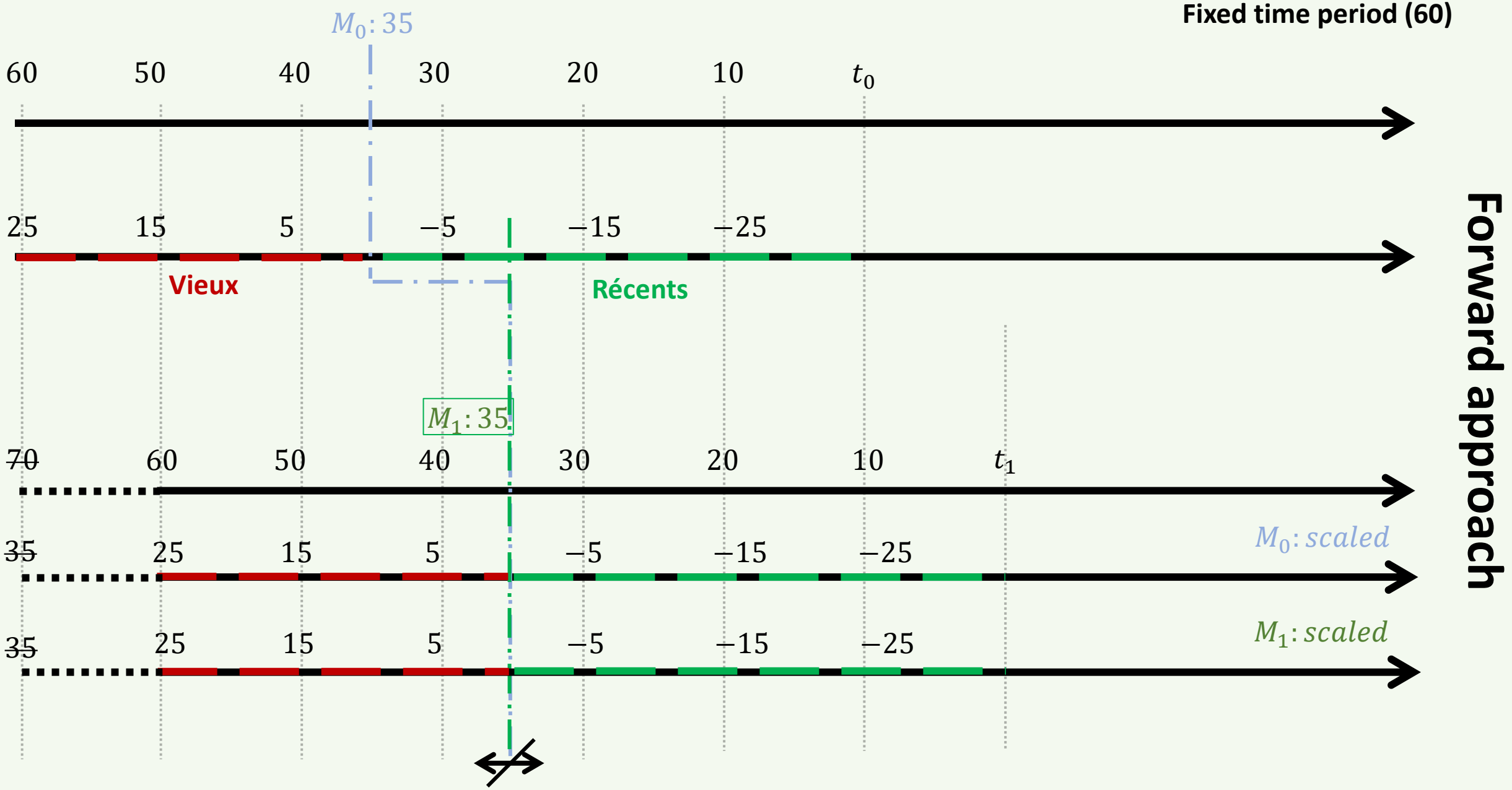

40
20
0
−20
−40
−40
−20
0
20
40

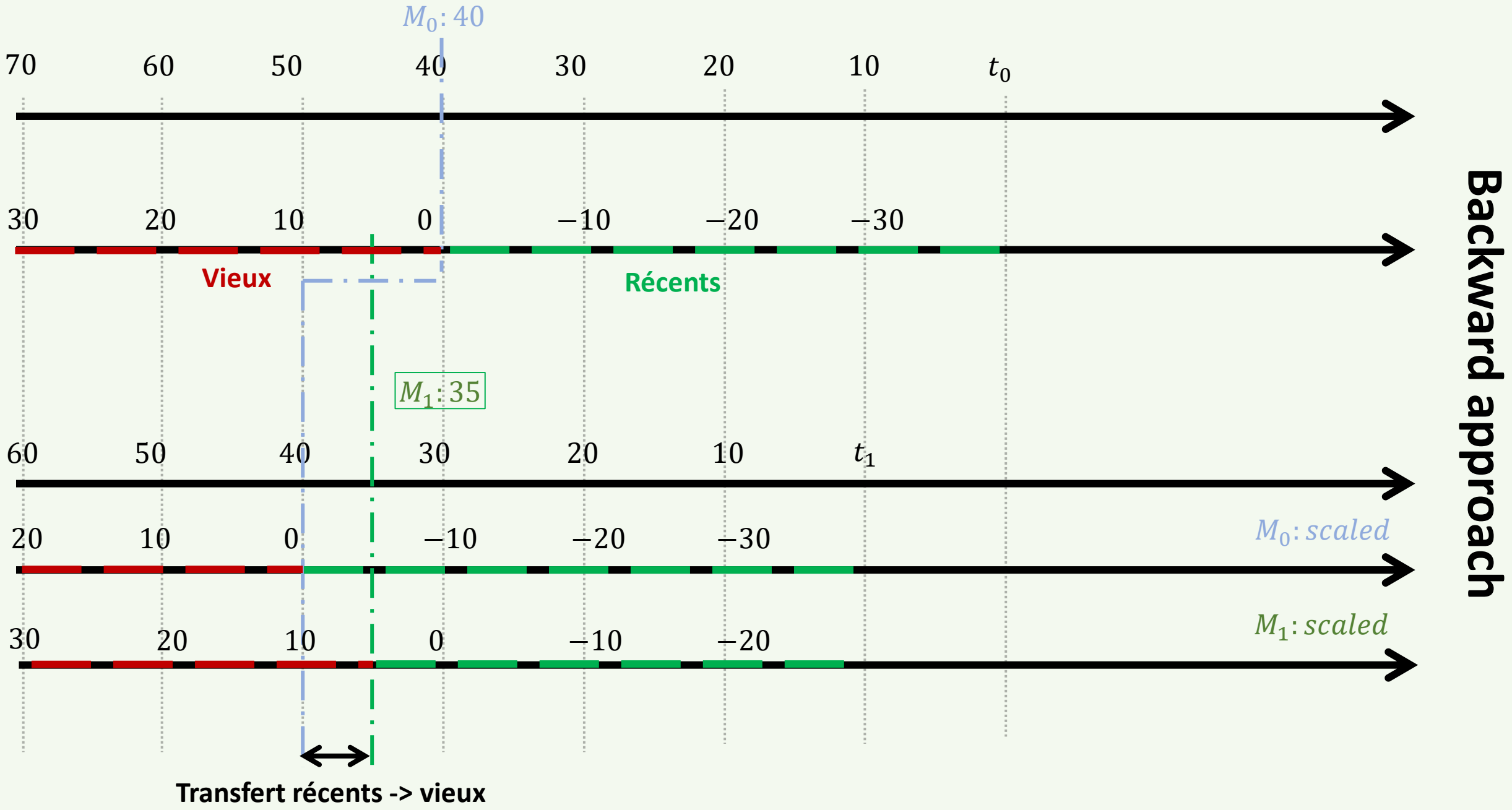

40
20
0
−20
−40
−40
−20
0
20
40

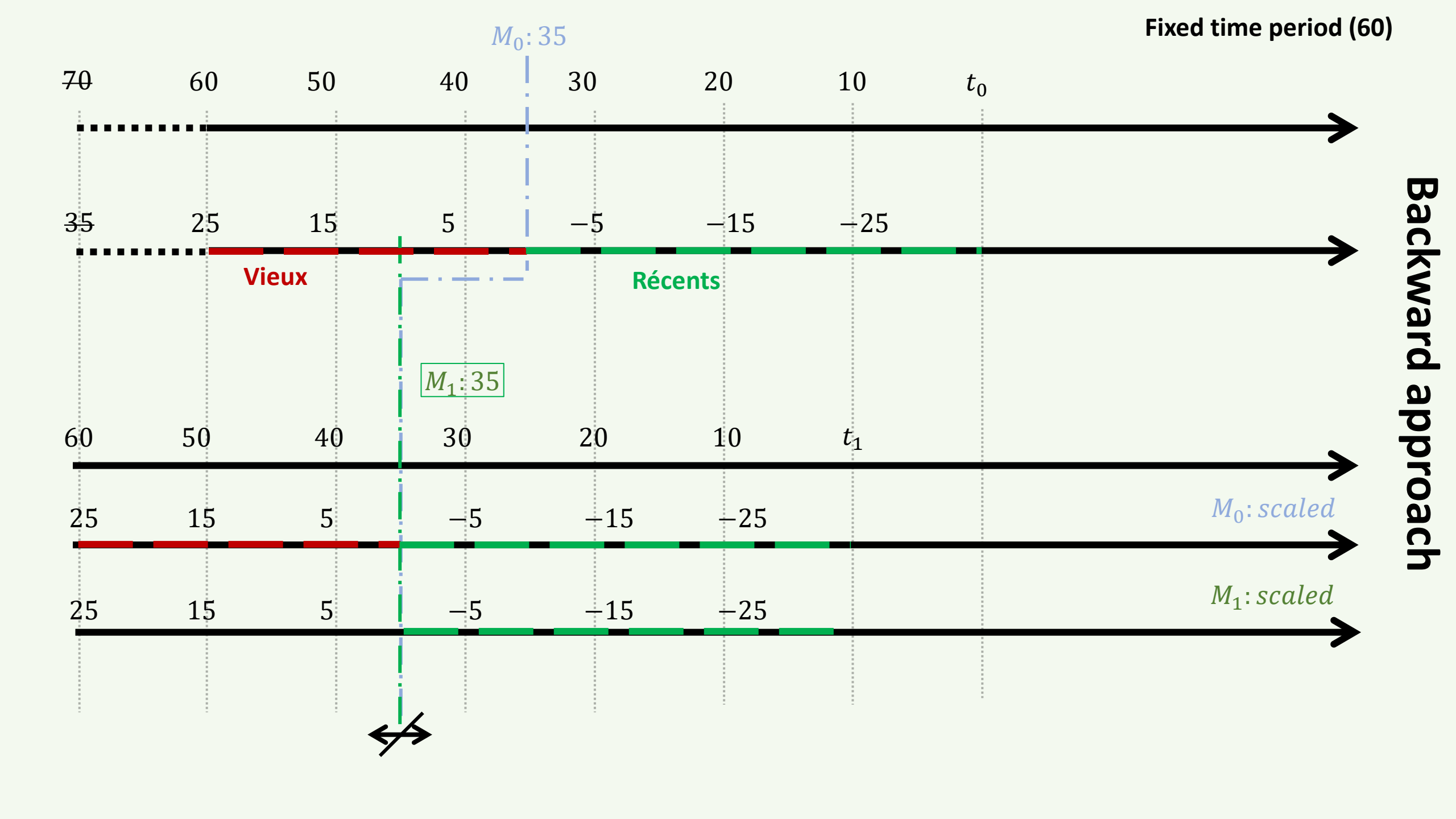

time
t_0
t_1
t_2
t_3
Modèle initial (#0) fit
Modèle initial (#0) prédiction
ARI / Acc. à t_1
Modèle (#1) fit & prédiction
Modèle initial (#0) prédiction
ARI / Acc. à t_2
Modèle (#2) fit & prédiction
...


$M_0: 35$
60
50
40
30
20
10
t_0
25
15
5
−5
−15
−25
Vieux
Récents
$M_1: 40$
70
60
50
40
30
20
10
t_1
35
25
15
5
−5
−15
−25
$M_0: scaled$
30
20
10
0
−10
−20
−30
$M_1: scaled$
Transfert vieux -> récents
Forward approach


$M_0: 35$
60 50 40 30 20 10 t_0
25 15 5 −5 −15 −25
Vieux
Récents
$M_1: 40$
70 60 50 40 30 20 10 t_1
$M_0: scaled$
$M_1: scaled$
$M_2: 45$
80 70 60 50 40 30 20 10 t_2
45 35 25 15 5 −5 −15 −25
$M_0: scaled$
35 25 15 5 −5 −15 −25 −35
$M_2: scaled$
Transfert vieux -> récents
Forward approach


Fixed time period (60)
$M_0: 35$
60
50
40
30
20
10
t_0
25
15
5
−5
−15
−25
Vieux
Récents
$M_1: 35$
70
60
50
40
30
20
10
t_1
35
25
15
5
−5
−15
−25
$M_0: scaled$
35
25
15
5
−5
−15
−25
$M_1: scaled$
Forward approach


$M_0: 40$
70
60
50
40
30
20
10
t_0
30
20
10
0
−10
−20
−30
Vieux
Récents
$M_1: 35$
60
50
40
30
20
10
t_1
20
10
0
−10
−20
−30
$M_0: scaled$
30
20
10
0
−10
−20
$M_1: scaled$
Transfert récents -> vieux
Backward approach


Fixed time period (60)
$M_0: 35$
70
60
50
40
30
20
10
t_0
35
25
15
5
−5
−15
−25
Vieux
Récents
$M_1: 35$
60
50
40
30
20
10
t_1
25
15
5
−5
−15
−25
$M_0: scaled$
25
15
5
−5
−15
−25
$M_1: scaled$
Backward approach

	Montant ++	Fréq ++	Pas content	Vieille	Récente
Montant ++	1297	11	0	0	0
Fréq ++	0	1923	0	0	0
Pas content	61	0	12318	0	0
Vieille	157	0	336	20451	2881
Récente	104	0	0	0	27736