

Apports et perspectives du Federated Learning dans le domaine assurantiel

Projet de Statistiques Appliquées

ENSAE Paris - Milliman France

21 Mai 2025

1. Introduction et Motivation
2. Données
3. Implémentation et Résultats - **FedAvg**
4. De nouvelles méthodes d'agrégation - **FedProx** et **FedOpt**
5. Implémentation et Résultats - **FedProx** et **FedOpt**
6. Conclusion

Objectif : prédire la sinistralité automobile à l'échelle européenne sans centraliser les données.

- Chaque assureur (France, Belgique, Espagne...) conserve ses propres données clients.
- La centralisation pose deux problèmes majeurs :
 - **Hypothèse i.i.d. non respectée** : clientèle hétérogène → distributions différentes entre compagnies.
Exemple : Certaines assurances automobiles ciblent des jeunes (Ornikar), d'autres des personnes plus âgées (Direct Assurance).
 - **Confidentialité** : RGPD, données sensibles (santé, infractions, géolocalisation...).

Solution proposée !

Utiliser l'**apprentissage fédéré** (*Federated Learning*) pour entraîner un modèle commun sans échange de données.

Qu'est-ce que le Federated Learning ?

- **Définition** : apprentissage collaboratif où plusieurs entités entraînent un modèle commun **sans partager leurs données locales**.

FL Vertical : entités détiennent des variables complémentaires sur les mêmes individus.

FL Horizontal : entités partagent des variables similaires mais sur des individus différents.

- **Exemples d'application** :

- *Santé* : hôpitaux entraînent un modèle de diagnostic sans échanger les dossiers patients.
- *Climat* : le modèle *IOFireNet* détecte les feux de forêt via capteurs/satellites sans compromettre la confidentialité.

Objectif : garantir une **performance homogène** sur l'ensemble des entités participantes.

Coefficients / Base de Données	A	B	C
A	✓✓	×	×
B	×	✓✓	×
C	×	×	✓✓
FL	✓	✓	✓

Table 1 – Performance prédictive croisée des modèles locaux vs. Federated Learning

Principe de l'Apprentissage Fédéré

Principes clés [Li et al., 2020]

- n_{round} cycles d'agrégation globale
- Entraînement local sur chaque client pendant E itérations correspondant aux poids locaux
- Mini-batches renouvelés à chaque round pour la diversité des données
- poids $w^{\text{agregated}}$ renvoyé à tous les clients pour reprendre un entraînement local
- Philosophie Open Weight : échange des poids des modèles et non des données !

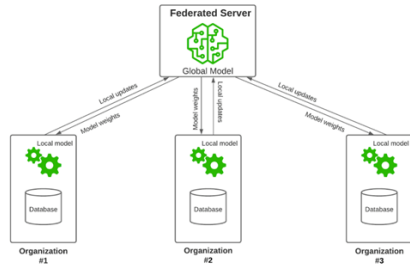


Figure 1 – Schéma général de l'apprentissage fédéré

Algorithm 1 Apprentissage Fédéré avec un mini-batch par round

```
1: Initialisation : le serveur choisit  $w_0$  (poids initiaux)
2: for  $t = 1$  à  $n_{\text{round}}$  do
3:   Le serveur diffuse  $w_{t-1}$  aux clients sélectionnés
4:   for chaque client  $i \in \{1, \dots, K\}$  (en parallèle) do
5:     Échantillonnage d'un mini-batch  $B_t^i$  local (différent à chaque round)
6:     for  $e = 1$  à  $E$  do ▷ Époques locales
7:        $w \leftarrow w - \alpha \nabla F_i(w; B_t^i)$    Avg et Opt, pour Prox
8:     end for
9:     Le client  $i$  retourne  $w_t^i$  au serveur
10:  end for
11:  Agrégation :  $w_t \leftarrow \text{Aggregate}(w_t^1, \dots, w_t^K)$ 
12: end for
```

Question

Quel modèle local est utilisé sur chaque client ?

Réponse : Régression Logistique.

- **Pourquoi ?** Modèle simple, robuste, interprétable — idéal pour des cas assurantiels.
- **Objectif** : prédire la probabilité d'un sinistre auto :
 $P(y_i^k = 1 \mid x_i^k, w_k) = \sigma((x_i^k)^T w_k)$.
- Chaque observation est associée à une Exposure $\in [0, 1]$.

Probabilité avec Exposure :

$$\begin{aligned} P(y_i^k = 1 \mid x_i^k, w_k) &= \sigma \left((x_i^k)^T w_k \right) \cdot f_i \\ &= \frac{1}{1 + e^{-(x_i^k)^T w_k}} \cdot f_i \end{aligned}$$

avec $f_i = \text{Exposure}_i \in [0, 1]$.

Formules du Modèle Local

1. Probabilité avec Exposure :

$$P(y_i^k = 1 \mid x_i^k, w_k) = \sigma \left((x_i^k)^T w_k \right) \cdot f_i = \frac{1}{1 + e^{-(x_i^k)^T w_k}} \cdot f_i$$

avec $f_i = \text{Exposure}_i \in [0, 1]$

2. Loi de Bernoulli conditionnelle :

$$P(y_i^k \mid x_i^k, w_k) = \left[\sigma((x_i^k)^T w_k) f_i \right]^{y_i^k} \left[1 - \sigma((x_i^k)^T w_k) f_i \right]^{1-y_i^k}$$

3. Vraisemblance (indépendance des obs.) :

$$\mathcal{L}(w_k) = \prod_{i=1}^{n_k} \left[\sigma((x_i^k)^T w_k) f_i \right]^{y_i^k} \left[1 - \sigma((x_i^k)^T w_k) f_i \right]^{1-y_i^k}$$

4. Fonction de coût (log-vraisemblance négative) utilisée par FedAvg et FedOpt :

$$F_k(w_k) = -\frac{1}{n_k} \sum_{i=1}^{n_k} \left[y_i^k \log \left(\sigma((x_i^k)^T w_k) f_i \right) + (1 - y_i^k) \log \left(1 - \sigma((x_i^k)^T w_k) f_i \right) \right]$$

Méthode d'Agrégation : FedAvg

Question

Niveau global, comment faire pour agréger les données ?

Réponse : Première méthode — FedAvg !

- **Principe** : moyenne pondérée des poids locaux w_k .
- **Pondération** : chaque client est pondéré par la taille n_k de son dataset.
- **Agrégation totale** : tous les clients participent à chaque itération.

$$w_{t+1}^{\text{FedAvg}} = \sum_{i=1}^K \frac{n_i}{\sum_j n_j} w_t^i$$

Problème

Dépendance de la taille entre bases, rapport de force inégale.

1. Introduction et Motivation
2. Données
3. Implémentation et Résultats - **FedAvg**
4. De nouvelles méthodes d'agrégation - **FedProx** et **FedOpt**
5. Implémentation et Résultats - **FedProx** et **FedOpt**
6. Conclusion

Objectif : Analyser les risques liés aux contrats d'assurance auto et aux sinistres.

Sources de données :

- **freMTPL (France)** : Fusion des 2 sous bases freMTPLfreq/sev 1 et 2.
- **beMTPL (Belgique)** : Données provenant de l'année 1997.
- **euMTPL (Europe, principalement Italie)** : données récolté sur 3 ans au début des années 2000 ; variable province = province italienne.

Variables clés :

- ClaimNb, ClaimAmount, Exposure
- Caractéristiques conducteur (âge, Sex) et véhicule (puissance, âge, carburant...)

Source : CASData Manual

Base	Nombre d'observations
freMTPL - France	1 091 182
beMTPL - Belgique	163 212
euMTPL - Europe/Italie	2 373 197

Table 2 – Quantité de données par base

Observation

La base européenne est prépondérante par rapport aux autres

Problème

Les 3 bases ne contiennent pas exactement les mêmes variables.

Objectif

Construire un ensemble cohérent de covariables $(\mathbf{x}_i^k)_{i=1}^{n_k}$.

Les transformations effectués :

- DriverAge : Variable commune, non modifiée.
- Sex : Absente dans freMTPL. Nécessité de la créer artificiellement
 - D'abord simuler aléatoirement (50% / 50%)
 - Puis réajuster selon :
 $p_h = 0.06$ (hommes sinistrés/hommes totaux), $p_f = 0.03$ (femmes sinistrées/femmes totaux)
- Objectif : introduire un *effet contrôlé* sur la sinistralité afin de tester la contribution dans les modèles prédictifs.

Variables reconstruites ou recodées (suite)

- **Power** : Codage différent dans la base française → harmonisation alphanumérique/numérique.
- **Fuel_type** : Codages hétérogènes entre les trois bases → recodée en binaire : Diesel / Regular.
- **Density** : Initialement uniquement dans freMTPL.
 - Estimée dans beMTPL et euMTPL via jointure avec les régions.
 - Données de densité récupérées via sources officielles belges et européennes.
- **Sinistre** :
 - Absente dans euMTPL, mais **ClaimAmount** est présente.
 - Transformation : si montant = 0 \Rightarrow pas de sinistre ; sinon \Rightarrow sinistre.

Remarque : Variable Sex n'est mobilisée qu'à des fins pédagogiques. Non utilisé en pratique dans les modèles de tarification (Voir *Délibération n° 2009-373 du 26 octobre 2009 relative aux discriminations dans l'accès à l'assurance* publiée par la HALDE).

Objectif

Comprendre la structure des données, détecter d'éventuelles anomalies et évaluer les distributions des variables avant d'appliquer toute méthode de modélisation

- Tous les graphiques ne seront pas analysés
- Matrice de corrélation, V de Cramer
- Toutes les statistiques descriptives sont disponibles en annexe

Conclusion : Statistique descriptive

Objectif : Comprendre la structure des données avant modélisation.

Principaux constats :

- DriverAge : Impact sur sinistralité variable selon les pays. Pas de tendance universelle.
- Sex : Effet non significatif sauf en France (effet artificiel simulé).
- Fuel_type : Légère corrélation avec sinistralité dans les bases belge et européenne.
- Power et Density : Variables très asymétriques, distributions différentes entre pays.
- Corrélation globale très faible entre toutes les variables continues et catégorielles.
- Sinistre : Données imbalanced.

Implication :

- Peu ou pas de multicolinéarité entre variables \Rightarrow conditions favorables à la régression.

Objectif

Garantir une échelle homogène entre les variables continues pour stabiliser la régression logistique avec régularisation.

- Méthode utilisée : **MinMaxScaler**

$$X_{\text{scaled}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- Variables concernées : DriverAge, Power, Density

Observation

Une minorité des assurés a déclaré au moins un sinistre :

- 4.5%, 7.7% et 11.2% selon les bases respectivement Française, European et Belge

Conséquences

- Données fortement déséquilibrées (classe 0 dominante)
- Risque d'un modèle biaisé vers l'absence de sinistre
- La précision peuvent se révéler trompeuses dans ce contexte.

Choix des métriques adaptées : **Courbe ROC (Receiver Operating Characteristic)**

Observation

Seuls 11.2% au maximum d'individus déclarent un sinistre sur l'ensemble des bases

Méthode utilisée : SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous)

- Génère des observations synthétiques pour la classe minoritaire
 - **Continue** : interpolation entre individus proches
 - **Catégorielle** : modalités choisies aléatoirement
-
- Appliqué uniquement sur l'échantillon d'**entraînement**, après séparation stratifiée
 - **Objectif** : améliorer l'apprentissage du modèle sur la classe minoritaire sans biaiser l'évaluation sur le test

Amélioration de la représentativité

- Rééquilibrage du ratio Sinistre à environ **30%** dans l'échantillon d'entraînement
- Meilleure capacité du modèle à apprendre la dynamique de la classe minoritaire

Stabilité des variables explicatives

- Distributions globalement conservées (âge, puissance, densité, etc.)
- Pas de déformation significative dans les données après augmentation

Test de significativité

- Tous les coefficients estimés pour les trois bases sont **hautement significatifs** ($p < 0,001$).
- Ceci confirme l'existence de relations statistiquement robustes entre les variables explicatives et la survenue d'un sinistre.
- Les détails chiffrés sont disponibles en annexe pour chaque base (France, Belgique, Europe).

Variable	Belgique (%)	Europe (%)	France (%)
Power	71,09	611,00	-56,27
DriverAge	-65,54	-27,79	13,66
Density	105,37	15,95	19,06
Sex (Homme)	-3,75	-4,74	94,88
Fuel_type (Diesel)	18,27	26,24	-3,26

Table 3 – Effets marginaux (en %) sur les odds de sinistre

- Variabilité importante selon les pays
- Ces divergences soulignent l'importance de conserver les spécificités locales dans la modélisation, d'où l'intérêt de l'apprentissage fédéré.

1. Introduction et Motivation
2. Données
3. Implémentation et Résultats - **FedAvg**
4. De nouvelles méthodes d'agrégation - **FedProx** et **FedOpt**
5. Implémentation et Résultats - **FedProx** et **FedOpt**
6. Conclusion

Paramétrage

- Nous avons utilisé une régression logistique avec une fonction de perte logarithmique, en intégrant un mécanisme de pondération automatique pour compenser le déséquilibre entre les classes (cost sensitive learning).
- Les poids estimés lors des itérations précédentes ont été réutilisés à chaque nouvelle phase d'entraînement dans le cadre de l'apprentissage fédéré, afin d'assurer une continuité dans l'optimisation du modèle.

Choix computationnels

Nous avons entraîné chaque mini-batch sur 60 % des données et testé sur 40 %. Une régularisation de type \mathcal{L}^2 a été appliquée pour limiter l'effet des variables trop dominantes, notamment l'intercept.

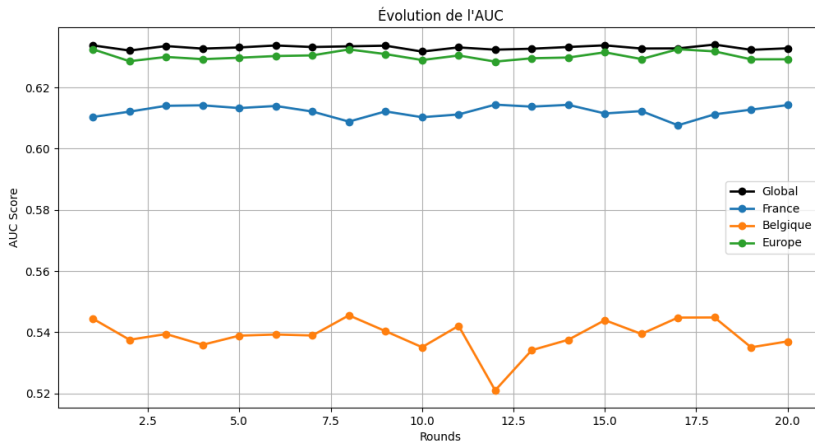


Figure 2 – Évolution du score AUC au fil des rounds pour FedAvg, par région

Résultats : Poids pour DriverAge

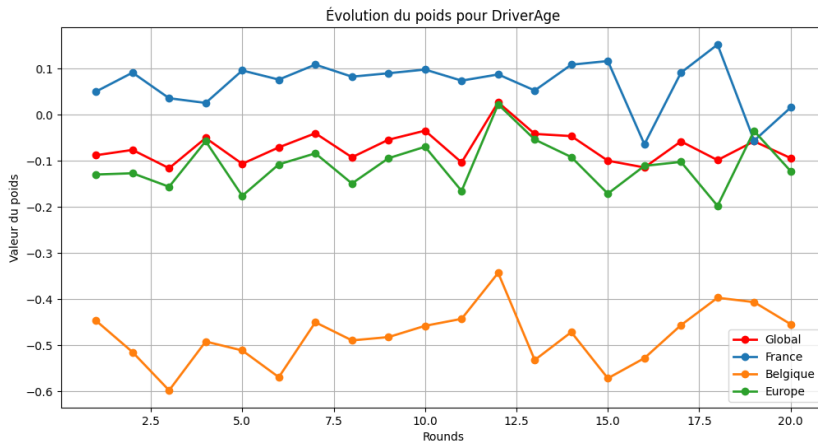


Figure 3 – Évolution des poids DriverAge par FedAvg

Résultats : Poids pour Density

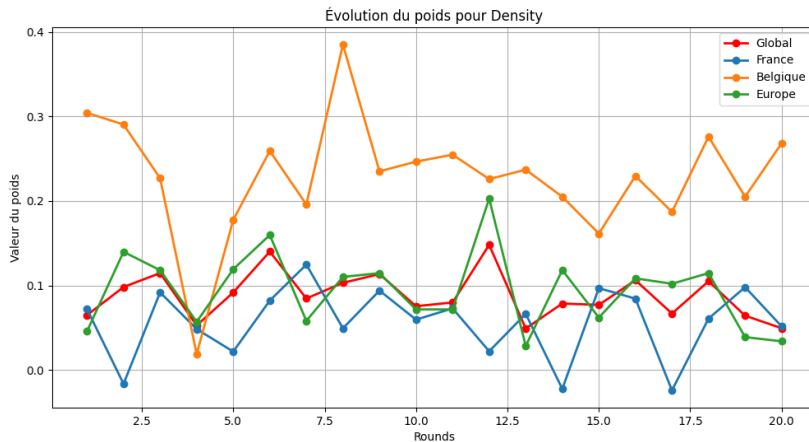


Figure 4 – Évolution des poids Density par FedAvg

- **Performance globale modeste** : le score AUC reste faible, inférieur à celui obtenu avec XGBoost.
- **Hétérogénéité inter-clients** : des disparités importantes entre les régions (France, Belgique, Europe) rendent l'agrégation difficile et nuisent à la performance du modèle global.
- **Instabilité de l'apprentissage** : la fonction de perte évolue de manière instable, notamment pour les régions belge et européenne, indiquant une convergence difficile.
- **Jeux de données déséquilibrés** : la petite taille de la base belge rend l'utilisation des mini-batches instable, ce qui affecte l'entraînement local.
- **Présence de biais dans certains coefficients** : certains résultats sont biaisés ou inattendus, comme le coefficient artificiellement élevé pour Sexe (France) ou très faible pour Power (Europe).
- **Performances inférieures aux modèles locaux** : sauf pour la base européenne, FedAvg n'améliore pas les résultats par rapport aux régressions logistiques locales.

1. Introduction et Motivation
2. Données
3. Implémentation et Résultats - **FedAvg**
4. De nouvelles méthodes d'agrégation - **FedProx** et **FedOpt**
 - 4.1 Federated Proximal - FedProx
 - 4.2 Federated Optimal - FedOpt
5. Implémentation et Résultats - **FedProx** et **FedOpt**
6. Conclusion

1. Introduction et Motivation
2. Données
3. Implémentation et Résultats - **FedAvg**
4. De nouvelles méthodes d'agrégation - **FedProx** et **FedOpt**
 - 4.1 Federated Proximal - FedProx
 - 4.2 Federated Optimal - FedOpt
5. Implémentation et Résultats - **FedProx** et **FedOpt**
6. Conclusion

Question

Comment mieux gérer l'hétérogénéité des données entre clients ?

Réponse : intégrer un terme de régularisation proximal !

- Même structure que FedAvg, mais fonction de coût modifiée.
- Le terme proximal limite la divergence locale.
- Idéal pour les données non-i.i.d.

Fonction de coût locale :

$$F_k^{prox}(w_k) = F_k(w_k) + \frac{\mu}{2} \|w_k - w^{FedAvg}\|_2^2$$

Mise à jour locale :

$$w_{t+1}^k = w_t^k - \alpha \nabla F_k(w_k, B) - \mu(w_t^k - w^{FedAvg})$$

1. Introduction et Motivation
2. Données
3. Implémentation et Résultats - **FedAvg**
4. De nouvelles méthodes d'agrégation - **FedProx** et **FedOpt**
 - 4.1 Federated Proximal - FedProx
 - 4.2 Federated Optimal - FedOpt
5. Implémentation et Résultats - **FedProx** et **FedOpt**
6. Conclusion

Question

Comment améliorer la convergence et la stabilité du modèle global avec données hétérogènes ?

Réponse : Optimisateurs adaptatifs côté serveur — FedOpt !

- Extension de FedAvg : même fonction de coût $F_k(w_k)$ par client, agrégation différente.
- Prise en compte de l'historique des gradients.
- Accumulateur de moments : stabilise les mises à jour via m_t .
- τ : terme de régularisation
- Δ_t : mise à jour moyenne pondérée des clients
- Mécanisme adaptatif de variance via u_t .

Moment de gradient :

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \Delta_t$$

Mise à jour globale :

$$w_{t+1}^{\text{FedOpt}} = w^{\text{FedOpt}} - \frac{\alpha}{\sqrt{u_t} + \tau} \cdot m_t$$

1. Introduction et Motivation
2. Données
3. Implémentation et Résultats - **FedAvg**
4. De nouvelles méthodes d'agrégation - **FedProx** et **FedOpt**
5. Implémentation et Résultats - **FedProx** et **FedOpt**
6. Conclusion

Implémentation

Nous conservons les mêmes hyperparamètres que dans l'implémentation de FedAvg.

Résultats

Les résultats obtenus avec FedProx montrent des performances globalement similaires à celles de FedAvg, sans réelle amélioration notable.

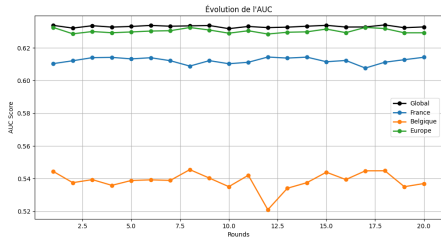


Figure 5 – AUC FedAvg

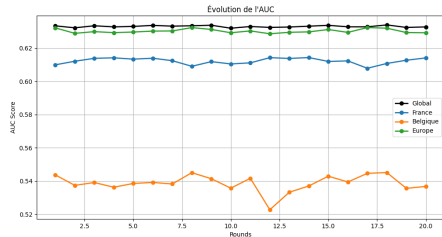


Figure 6 – AUC FedProx

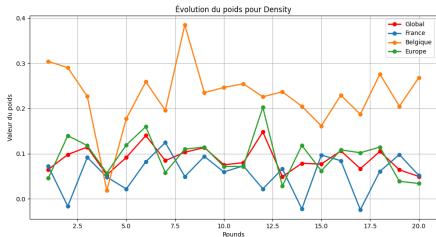


Figure 7 – Density FedAvg

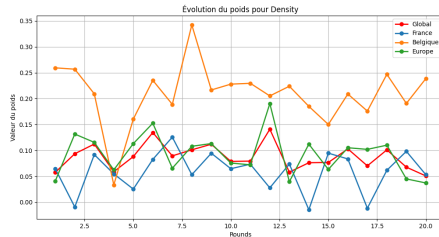


Figure 8 – Density FedProx

Implémentation et résultats - FedOpt

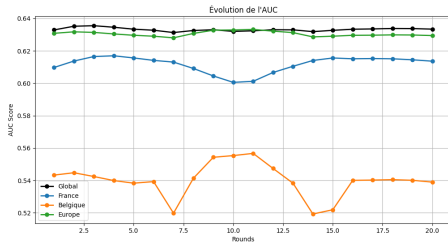


Figure 9 – Évolution du score AUC au fil des rounds pour FedOpt, par région

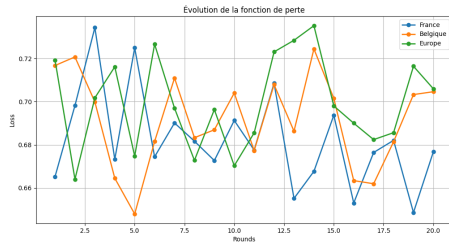


Figure 10 – Évolution de la fonction de perte au fil des rounds pour FedOpt

Implémentation et résultats - FedOpt

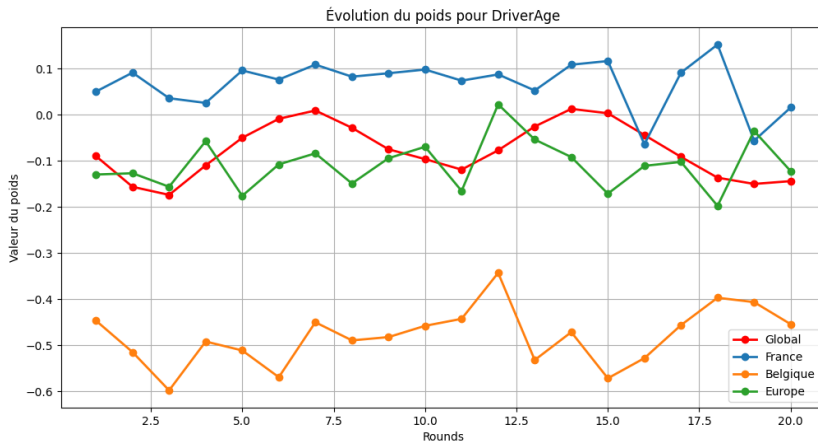



Figure 11 – Évolution des poids DriverAge par FedOpt (Adam)


Source	Données Test	AUC Local	AUC FedOpt	Amélioration(%)
France	France	0.6242	0.6098	-1.44%
France	Belgique	0.5063	0.5455	+3.92%
France	Europe	0.6201	0.6319	+1.18%
Belgique	France	0.5968	0.6098	+1.31%
Belgique	Belgique	0.5818	0.5455	-3.63%
Belgique	Europe	0.6379	0.6319	-0.60%
Europe	France	0.5980	0.6098	+1.18%
Europe	Belgique	0.5669	0.5455	-2.14%
Europe	Europe	0.6371	0.6319	-0.53%


Table 4 – Évaluation croisée FedOpt pour chaque région source et cible.

1. Introduction et Motivation
2. Données
3. Implémentation et Résultats - **FedAvg**
4. De nouvelles méthodes d'agrégation - **FedProx** et **FedOpt**
5. Implémentation et Résultats - **FedProx** et **FedOpt**
6. Conclusion

- Les méthodes d'**agrégation** jouent un rôle central : *FedOpt* a offert les meilleures performances, au prix d'une moindre interprétabilité.
- L'**hétérogénéité inter-clients** (données, expositions, tailles) affecte fortement les résultats globaux.
- Les bases de données limitées (ex. : Belgique) nuisent à la robustesse de la fédération.
- Les stratégies d'**augmentation de données** comme SMOTENC ont des effets contrastés ; l'*under-sampling contrôlé* pourrait être une alternative plus naturelle.
- Des questions clés restent ouvertes : *robustesse, équité, pondérations adaptatives, régularisation locale...*

 Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020).
A unified linear speedup analysis of federated averaging and nesterov fedavg.
arXiv preprint arXiv :2007.05690.

 Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konecný, J., Kumar, S., and McMahan, H. B. (2020).
Adaptive federated optimization.
arXiv preprint arXiv :2003.00295.

 Yang, N., Xin, Y., Lin, H., Lyu, P., and Wang, J. (2024).
FedADM : Adaptive federated learning via dissimilarity measure.

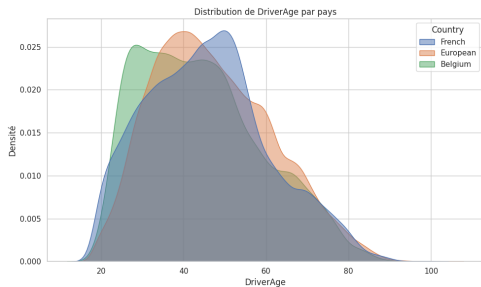


Merci !

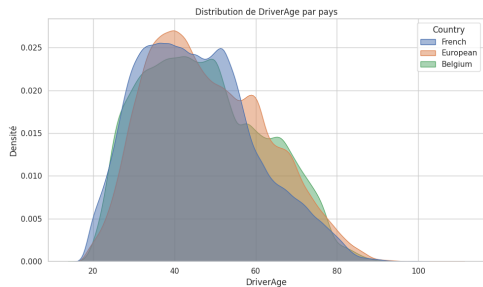


INSTITUT
POLYTECHNIQUE
DE PARIS

Annexe - Statistiques descriptives



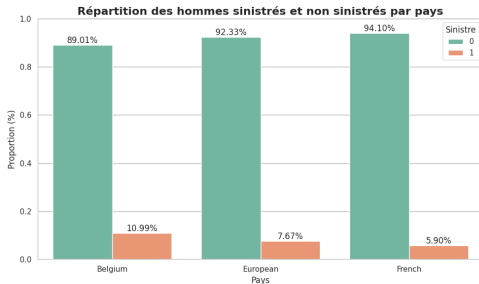
(a) Distribution de l'âge des conducteurs parmi les sinistres



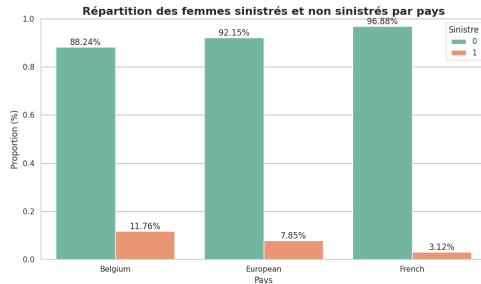
(b) Distribution de l'âge des conducteurs parmi les non sinistres

Figure 12 – Visualisation de la distribution de l'âge des conducteurs

Annexe - Statistiques descriptives (suite)



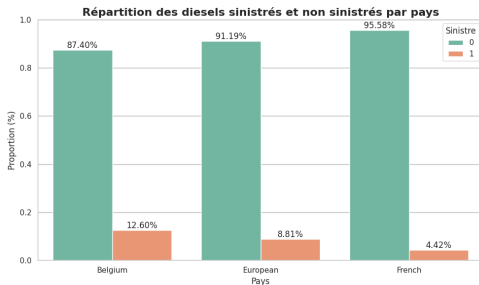
(a) Proportion d'homme sinistré et non sinistré



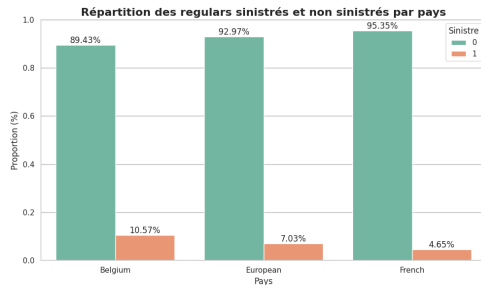
(b) Proportion de femme sinistrée et non sinistré

Figure 13 – Visualisation de la distribution du Sex des conducteurs

Annexe - Statistiques descriptives (suite)



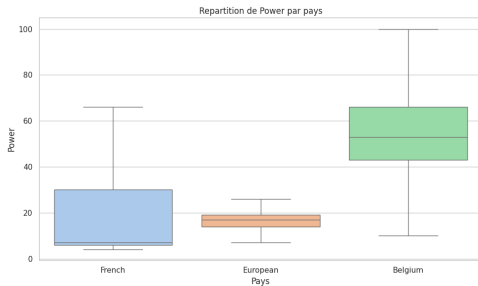
(a) Distribution de Diesel sinistré et non sinistré



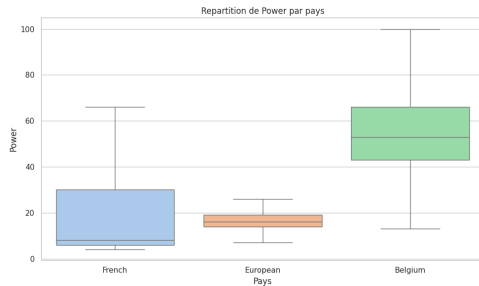
(b) Distribution de Regular sinistré et non sinistré

Figure 14 – Visualisation de la distribution du carburant des conducteurs

Annexe - Statistiques descriptives (suite)



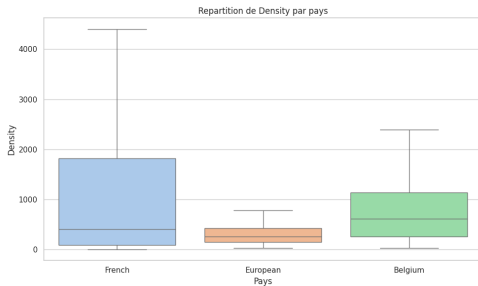
(a) Boxplot de la puissance du véhicule des sinistres



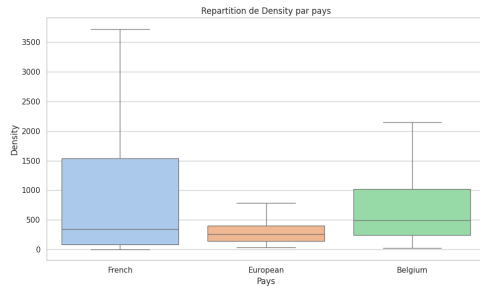
(b) Boxplot de la puissance du véhicule des non sinistres

Figure 15 – Visualisation de la distribution de la puissance du véhicule

Annexe - Statistiques descriptives (suite)



(a) Boxplot de la densité de population des sinistres

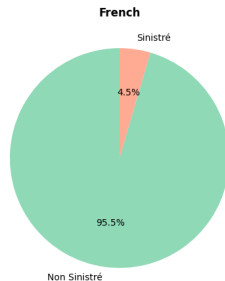
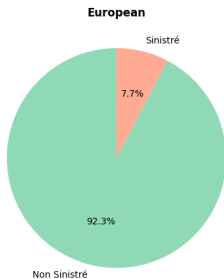
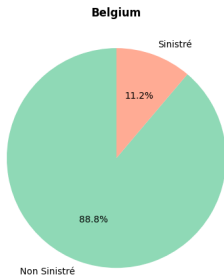


(b) Boxplot de la densité de population des sinistres

Figure 16 – Visualisation de la distribution de la densité de population des sinistres

Annexe — Regressions logistiques

Distribution des Sinistres par Pays



Annexe — Régressions logistiques(Base Belge et Européenne)

Variable	Coef	Std. Err.	z	P> z	[0.025, 0.975]
Constante	-0.6425	0.017	-37.625	0.000	[-0.676, -0.609]
Power	0.5369	0.060	8.941	0.000	[0.419, 0.655]
DriverAge	-1.0614	0.026	-40.649	0.000	[-1.113, -1.010]
Fuel_type	0.1678	0.010	16.062	0.000	[0.147, 0.188]
Density	0.7202	0.032	22.724	0.000	[0.658, 0.782]
Sex	-0.0382	0.011	-3.418	0.001	[-0.060, -0.016]

Table 5 – Résultats de la régression logistique sur la base belge

Constante	-0.8672	0.006	-149.357	0.000	[-0.879, -0.856]
Power	1.9634	0.325	6.045	0.000	[1.327, 2.600]
DriverAge	-0.3254	0.008	-40.563	0.000	[-0.341, -0.310]
Fuel_type	0.2330	0.003	81.959	0.000	[0.227, 0.239]
Density	0.1480	0.005	30.865	0.000	[0.139, 0.157]
Sex	-0.0486	0.003	-18.389	0.000	[-0.054, -0.043]

Table 6 – Résultats de la régression logistique sur la base européenne

Variable	Coef	Std. Err.	z	P> z	[0.025, 0.975]
Constante	-1.1779	0.005	-225.763	0.000	[-1.188, -1.168]
Power	-0.8288	0.012	-66.505	0.000	[-0.853, -0.804]
DriverAge	0.1281	0.010	12.389	0.000	[0.108, 0.148]
Fuel_type	-0.0331	0.004	-9.034	0.000	[-0.040, -0.026]
Density	0.1745	0.011	15.658	0.000	[0.153, 0.196]
Sex	0.6671	0.004	179.147	0.000	[0.660, 0.674]

Table 7 – Résultats de la régression logistique sur la base française

Statistiques descriptives après data augmentation

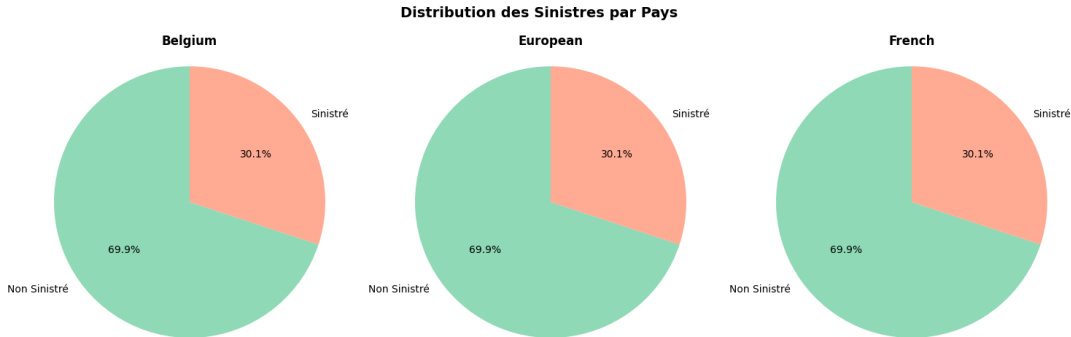
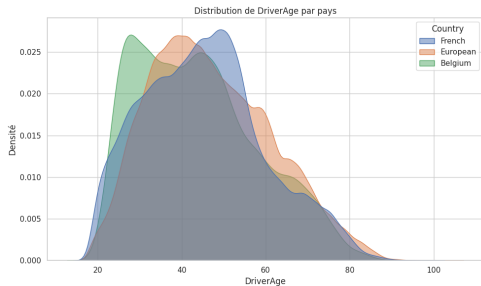
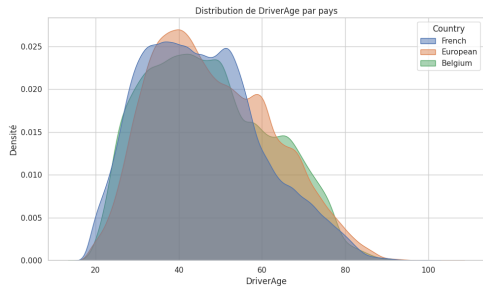


Figure 17 – Distribution des Sinistres après data augmentation

Statistiques descriptives après data augmentation(Suite)



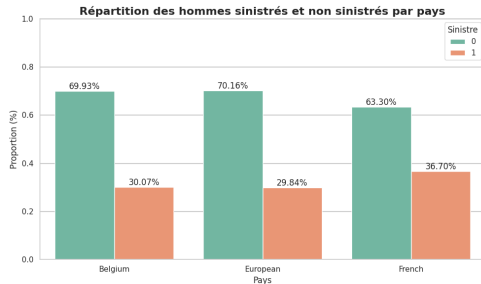
(a) Distribution de l'âge des conducteurs parmi les sinistres



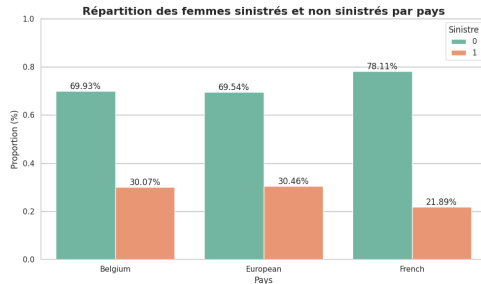
(b) Distribution de l'âge des conducteurs parmi les nons sinistres

Figure 18 – Visualisation de la distribution de l'âge des conducteurs après data augmentation

Statistiques descriptives après data augmentation(Suite)



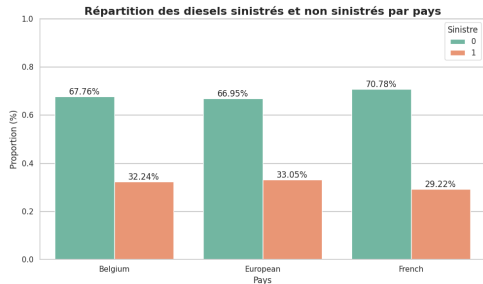
(a) Distribution de la proportion d'homme sinistré et non sinistré



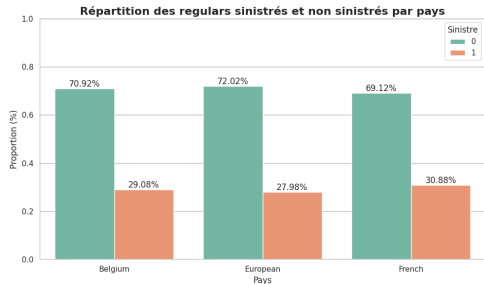
(b) Distribution de la proportion de femme sinistrée et non sinistrée

Figure 19 – Visualisation de la distribution du Sex des conducteurs après data augmentation

Statistiques descriptives après data augmentation(Suite)



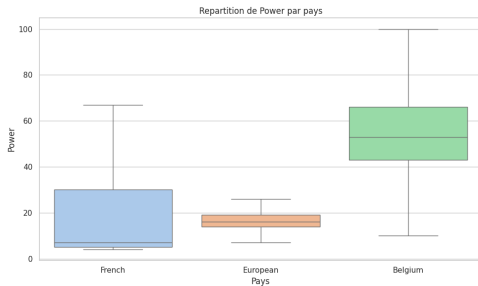
(a) Distribution du carburant Diesel selon les sinistres



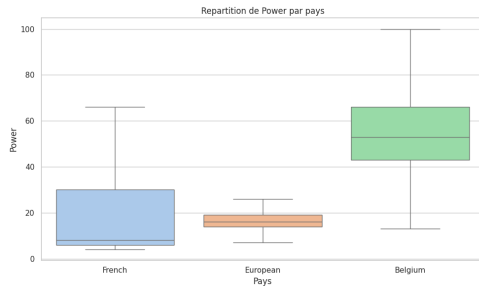
(b) Distribution du carburant Regular selon les sinistres

Figure 20 – Visualisation de la distribution du carburant des conducteurs après data augmentation

Statistiques descriptives après data augmentation(Suite)



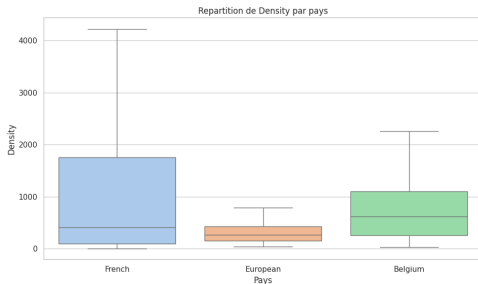
(a) Boxplot de la puissance du véhicule des sinistres



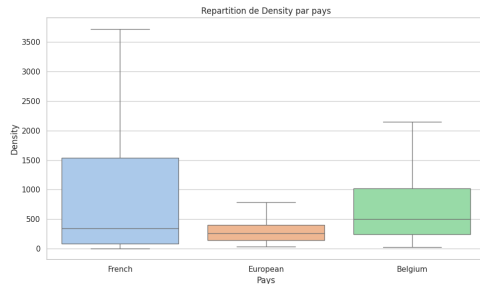
(b) Boxplot de la puissance du véhicule des non sinistres

Figure 21 – Visualisation de la distribution de la puissance du véhicule après data augmentation

Statistiques descriptives après data augmentation(Suite)



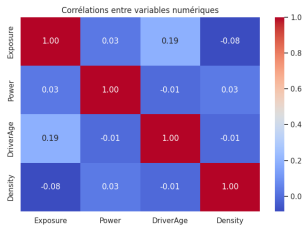
(a) Boxplot de la densité de population des sinistres



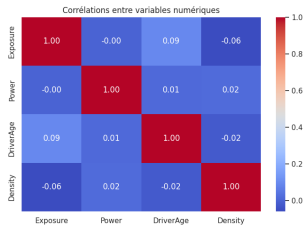
(b) Boxplot de la densité de population des sinistres

Figure 22 – Visualisation de la distribution de la densité de population des sinistres après data augmentation

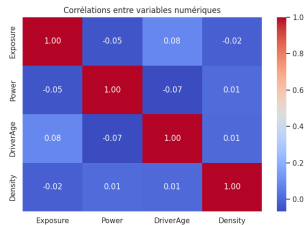
Statistiques descriptives après data augmentation(Suite)



(a) France



(b) Belge



(c) Européenne

Figure 23 – Matrice de corrélation de chaque base après data augmentation

Statistiques descriptives après data augmentation(Suite)

Table 8 – Résultats de l'analyse du V de Cramér entre variables catégorielles après data augmentation

Données	Variables	Cramer's V	Corrélation	p-value	Significatif
French	Fuel type - Sinistre	0.018	Faible	0.000	Oui
French	Fuel type - Sex	0.007	Faible	0.000	Oui
French	Sinistre - Sex	0.161	Modérée	0.000	Oui
Belgium	Fuel type - Sinistre	0.032	Faible	0.000	Oui
Belgium	Fuel type - Sex	0.109	Modérée	0.000	Oui
Belgium	Sinistre - Sex	0.000	Faible	0.098	Non
European	Fuel type - Sinistre	0.054	Faible	0.000	Oui
European	Fuel type - Sex	0.151	Modérée	0.000	Oui
European	Sinistre - Sex	0.007	Faible	0.000	Oui

Tableau récapitulatif des variables

Variable	Description	Codage
DriverAge	Âge du conducteur principal	Entier (ex. : 45)
Sex	Sexe du conducteur	Binaire : Male = 1, Female = 0
Power	Puissance du véhicule (en kW)	Entier (ex. : 85)
Fuel_type	Type de carburant	Binaire : Diesel = 1, Regular = 0
Density	Densité de population de la zone de résidence	Entier (ex. : 1200 hab/ km ²)
Sinistre	Présence d'un sinistre	Binaire : Oui = 1, Non = 0

Table 9 – Résumé des variables et de leur codage

Question

Pourquoi le XGBoost ?

Réponse : Dans notre contexte de classes déséquilibrées, la précision n'étant pas fiable, nous avons privilégié des métriques robustes comme l'AUC pour évaluer les performances. Pour disposer d'une borne supérieure théorique, nous avons entraîné un modèle XGBoost, et comparé ses résultats à ceux du modèle fédéré ainsi qu'à des régressions logistiques locales.

Annexe Comparaison AUC XGBoost et AUC Régression logistique

Base	AUC XGBoost	AUC Régression Logistique
freMTPL - France	0.6729	0.55
beMTPL - Belgique	0.5987	0.56
euMTPL - Europe/Italie	0.6605	0.57

Table 10 – AUC obtenus avec XGBoost et régression logistique locale avant data augmentation

Base	AUC XGBoost	AUC Régression Logistique
freMTPL - France	0.7200	0.6423
beMTPL - Belgique	0.7059	0.5907
euMTPL - Europe/Italie	0.6709	0.6436

Table 11 – AUC obtenus avec XGBoost et régression logistique locale après data augmentation

Annexe — Poids des variables dans FedAvg

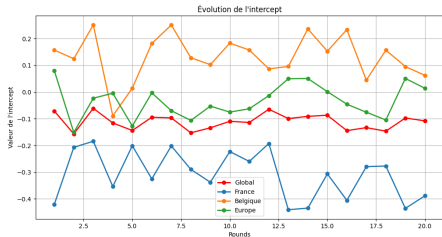


Figure 24 – Intercept FedAvg

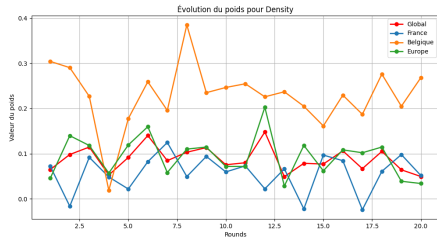


Figure 25 – Density FedAvg

Annexe — Poids des variables dans FedAvg

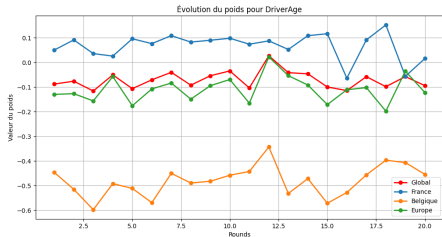


Figure 26 – DriverAge FedAvg

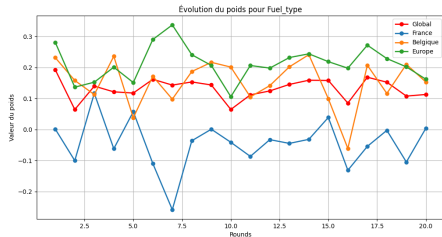


Figure 27 – Fueltype FedAvg

Annexe — Poids des variables dans FedAvg

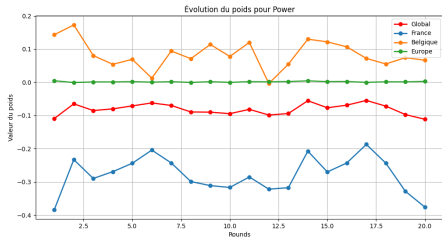


Figure 28 – Power FedAvg

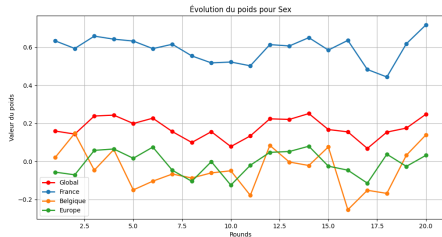


Figure 29 – Sex FedAvg

Règle de mise à jour locale (itération t) :

$$w_{t+1}^k = w_t^k - \alpha \nabla F_k(w_k, B) - \mu(w_t^k - w^{\text{FedAvg}})$$

Extension : ajustement dynamique de $\mu_{t,k}$ (FedADM) [Yang et al., 2024]

$$\mu_{t,k} = \mu_{t-1,k} + \alpha (\|w_k - w_{t,\text{Global}}\| - \xi)$$

Logique :

- Si $\|w_k - w_{t,\text{Global}}\| > \xi$: $\mu_{t,k} \uparrow \rightarrow$ plus de régularisation.
- Si $\|w_k - w_{t,\text{Global}}\| < \xi$: $\mu_{t,k} \downarrow \rightarrow$ plus de flexibilité locale.

Remarque : ce mécanisme n'a pas été implémenté par manque de temps. Par défaut, $\mu = 0.1$.

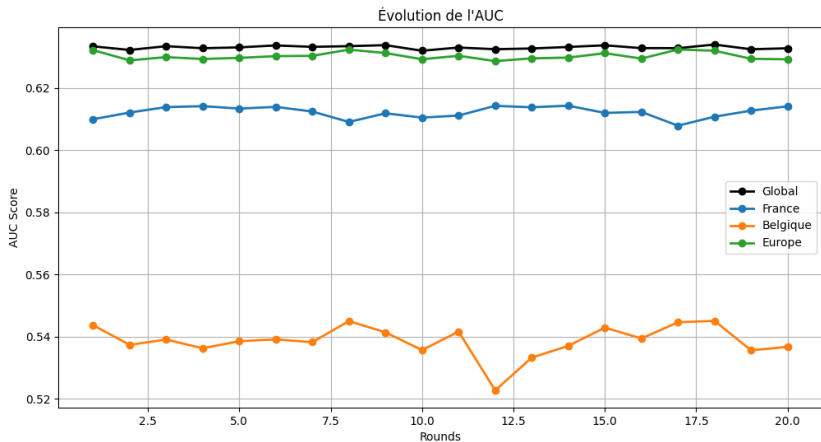


Figure 30 – Évolution du score AUC au fil des rounds pour FedProx, par région

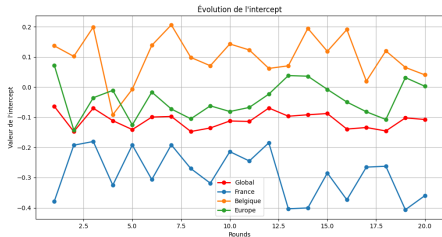


Figure 31 – Intercept FedProx

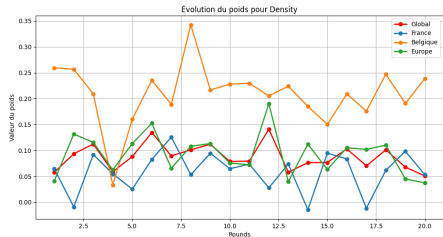


Figure 32 – Density FedProx

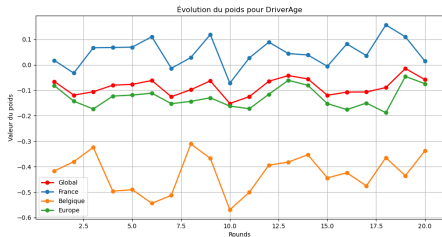


Figure 33 – DriverAge FedProx

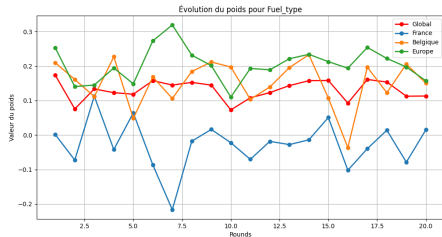


Figure 34 – Fuel_type FedProx

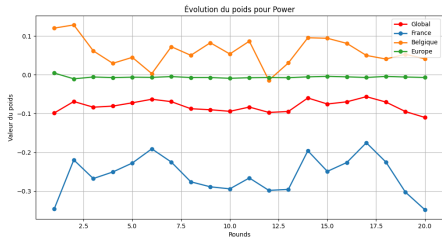


Figure 35 – Power FedProx

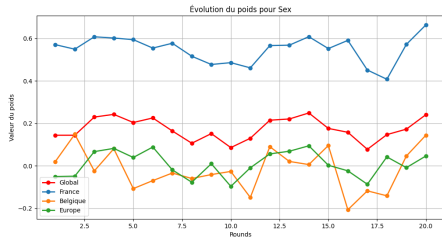


Figure 36 – Sex FedProx

Estimation de la variance adaptative u_t :

- **FedAdagrad** : $u_t = u_{t-1} + (\Delta_t)^2$
- **FedAdam** : $u_t = \beta_2 u_{t-1} + (1 - \beta_2)(\Delta_t)^2$
- **FedYogi** : $u_t = u_{t-1} - (1 - \beta_2) \cdot \text{sign}(u_{t-1} - (\Delta_t)^2) \cdot (\Delta_t)^2$

Recommandations pratiques [Reddi et al., 2020] :

- $\beta_1 = 0.9, \beta_2 = 0.99$
- $\tau = 10^{-3}$ pour stabiliser la division

Les paramètres sont choisis pour minimiser la perte d'entraînement moyenne sur les 100 dernières itérations.

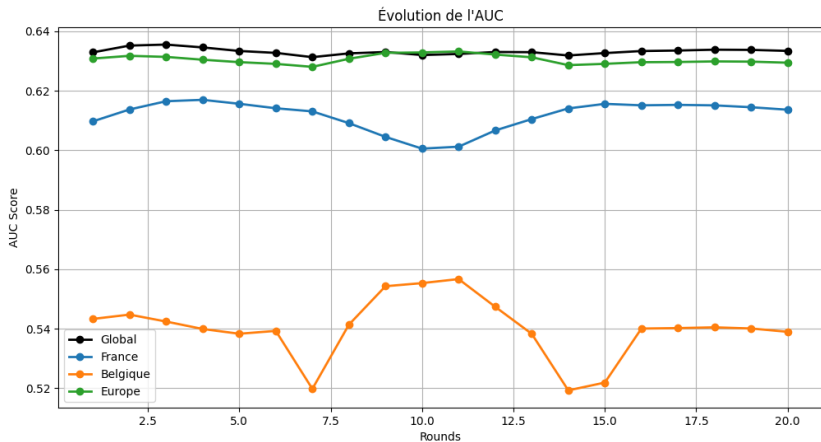


Figure 37 – Évolution du score AUC au fil des rounds pour FedOpt, par région

Annexe — Résultats FedOpt

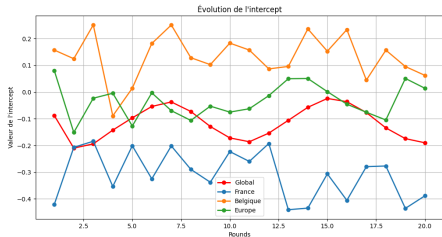


Figure 38 – Intercept FedOpt

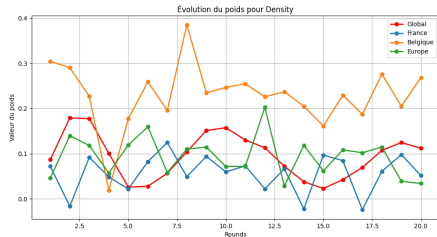


Figure 39 – Density FedOpt

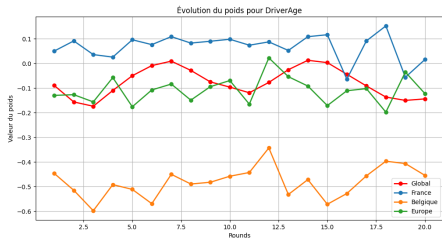


Figure 40 – DriverAge FedOpt

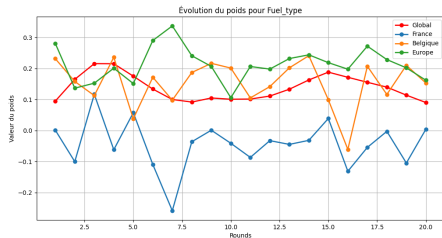


Figure 41 – Fueltype FedOpt

Annexe — Résultats FedOpt

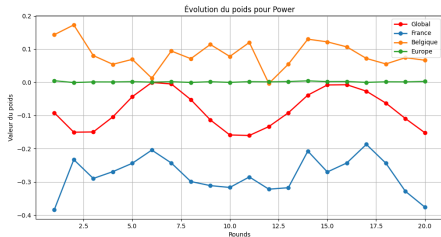


Figure 42 – Power FedOpt

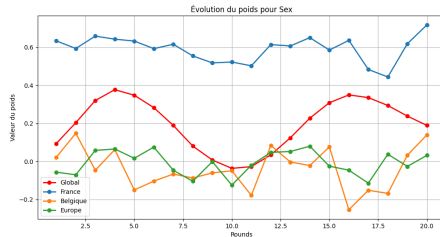


Figure 43 – Sex FedOpt