

Prédiction séquence-à-séquence de séries temporelles multivariées et déséquilibrées avec réseaux de neurones convolutifs unidimensionnels

Mercredi 27 Janvier 2021



Mehdi Elion



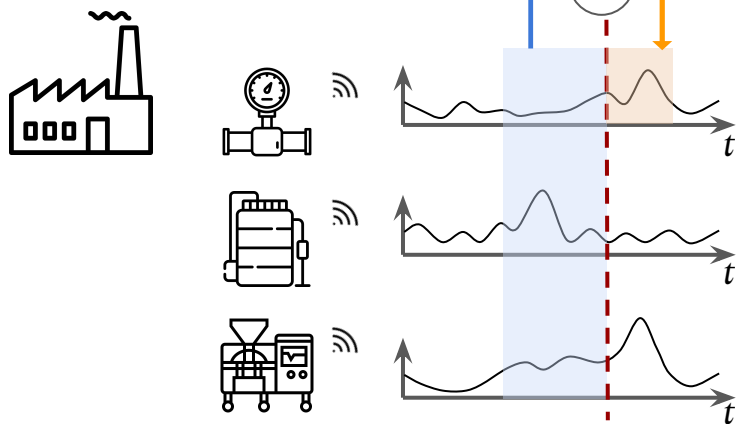
Sonia Tabti



Julien Budynek

Contexte et motivation

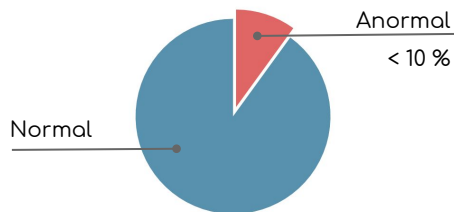
Données industrielles



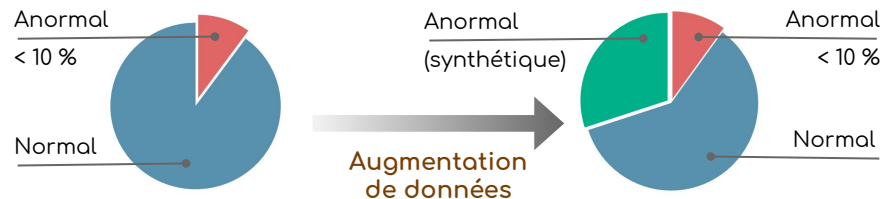
Évènements rares

- Pannes
- Anomalies, non conformités
- Baisses de qualité

Données déséquilibrées



Augmentation de données

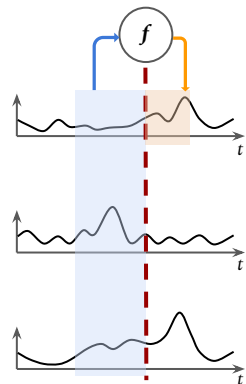
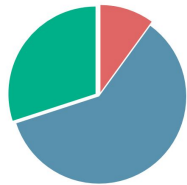


État de l'art

2002	• SMOTE (Synthetic Minority Oversampling Technique for classification)	Chawla et al.
2013	• SMOTE for Regression	Torgo et al.
2017	• Adaptation de SMOTE-R à la prédiction de séries temporelles univariées à un pas dans le futur	Moniz et al.
2020	• SMOTEST (Synthetic Minority Oversampling Technique for Sequence-To-sequence)	Notre proposition

Approche proposée : SMOTEST

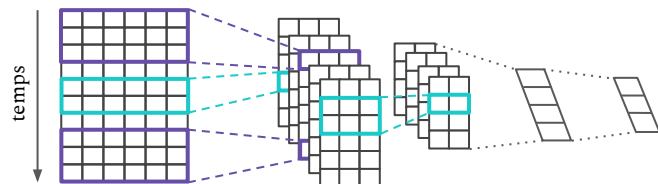
Sur-échantillonnage des données d'entraînement



Séries temporelles multivariées

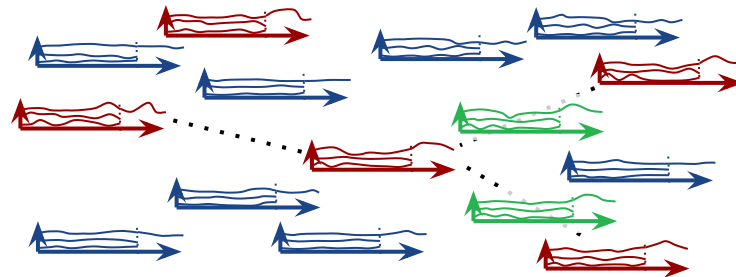
Séquence-à-séquence

CNN-1D

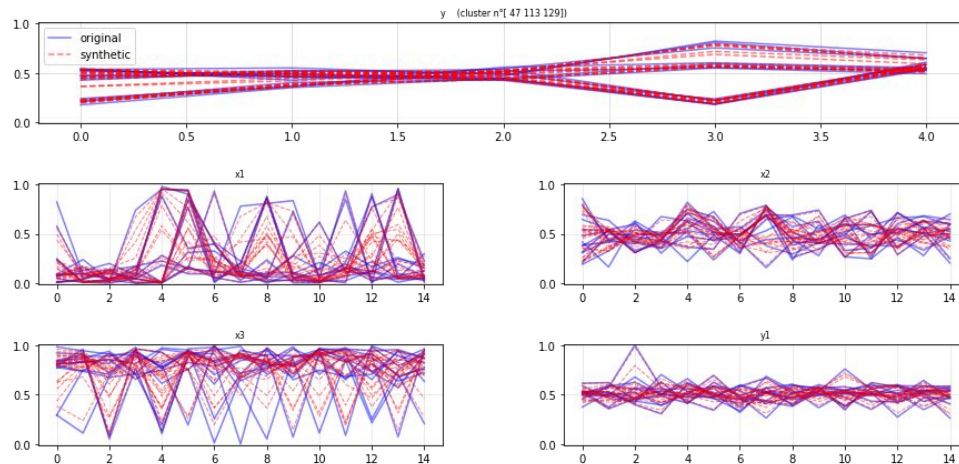


Algorithme de sur-échantillonnage

Illustration

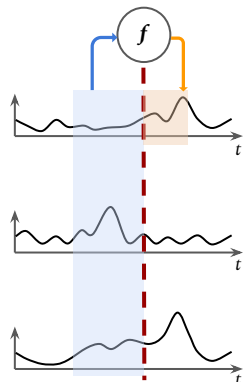
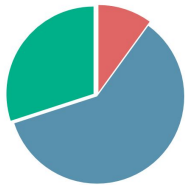


Exemple sur jeu de données synthétiques



Approche proposée : SMOTEST

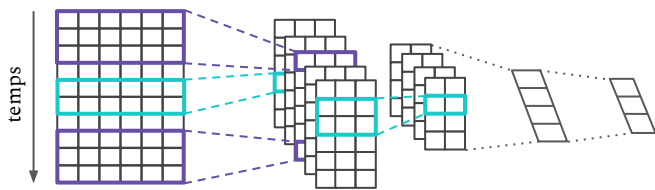
Sur-échantillonnage des données d'entraînement



Séries temporelles multivariées

Séquence-à-séquence

CNN-1D



Algorithme de sur-échantillonnage

Algorithm 1 Algorithme de sur-échantillonnage

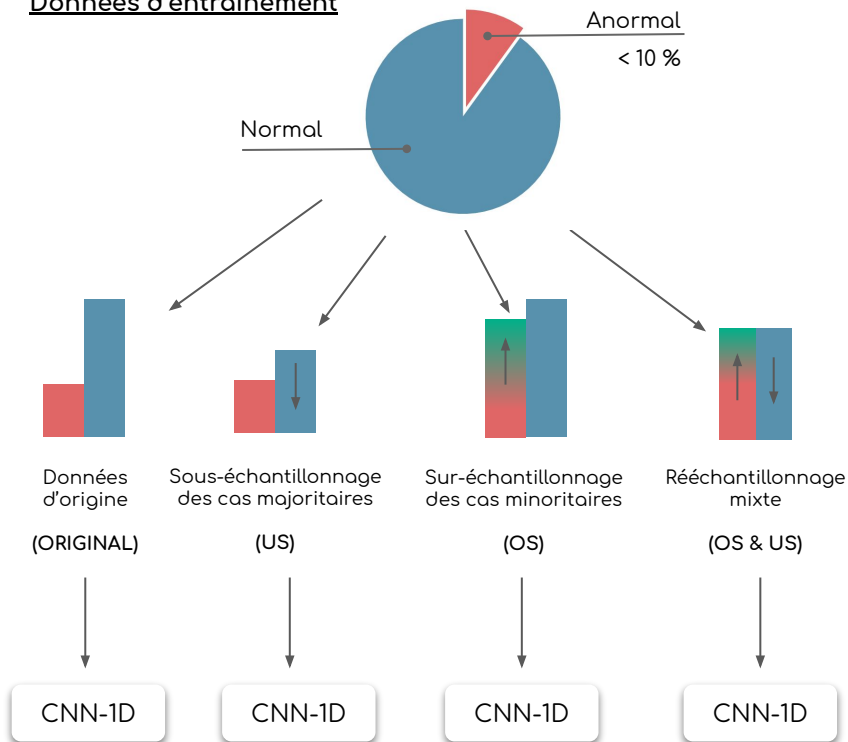
Require: n_g : nombre de cas synthétiques à générer par cas positif existant, k : nombre de voisins, G : générateur de nombres aléatoires, $\mathcal{D}_P = \{(X, Y) \mid \Phi(Y) \geq \theta\}$: ensemble des cas positifs

```
1: function GENSYNTHCASES( $\mathcal{D}_P, n_g, k, G$ )
2:    $\mathcal{D}_{gen} \leftarrow \emptyset$ 
3:   for all  $(X, Y) \in \mathcal{D}_P$  do
4:      $NNS \leftarrow KNN(k, Y, \mathcal{D}_P \setminus \{(X, Y)\})$ 
5:     for  $i = 1$  to  $n_g$  do
6:        $(X_{nn}, Y_{nn}) \leftarrow$  un cas choisi aléatoirement parmi  $NNS$ 
7:       Initialiser  $(X_{new}, Y_{new})$  pour contenir le  $i^{\text{ème}}$  cas synthétique
8:       for all  $s \in [1, n_X]$  do
9:          $diff \leftarrow X_{nn}[:, s] - X[:, s]$ 
10:         $X_{new}[:, s] \leftarrow X[:, s] + G(0, 1) \times diff$ 
11:       end for
12:        $d_1 \leftarrow DIST(X_{new}, X)$ 
13:        $d_2 \leftarrow DIST(X_{new}, X_{nn})$ 
14:        $Y_{new} \leftarrow \frac{d_2 Y + d_1 Y_{nn}}{d_1 + d_2}$ 
15:        $\mathcal{D}_{gen} \leftarrow \mathcal{D}_{gen} \cup \{(X_{new}, Y_{new})\}$ 
16:     end for
17:   end for
18:   return  $\mathcal{D}_{gen}$ 
19: end function
```

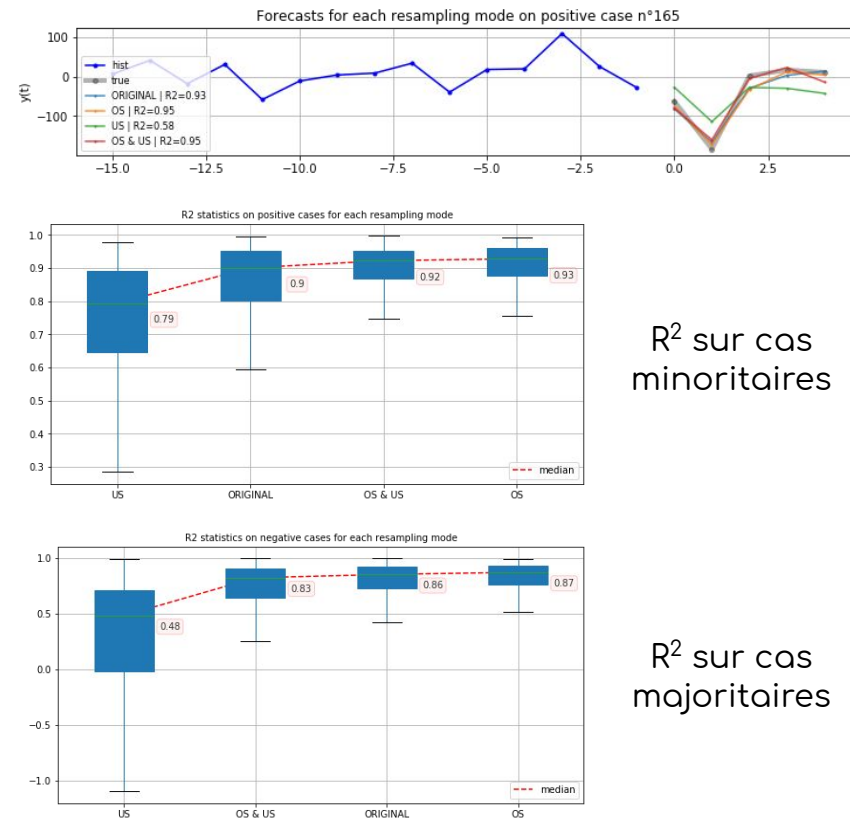
▷ distance euclidienne

Influence du mode de rééchantillonnage

Données d'entraînement



Résultats sur données synthétiques



R^2 sur cas minoritaires

R^2 sur cas majoritaires

Influence de la distribution aléatoire

Fonctions forme

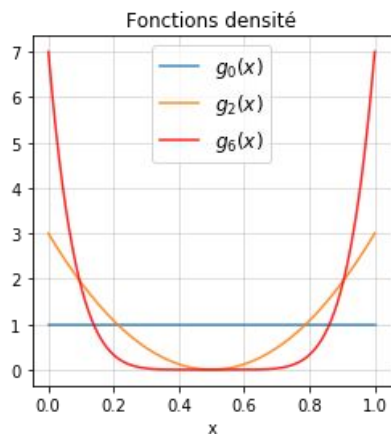
$$f_n(x) = (x - 0.5)^n$$

Fonctions densité

$$g_n(x) = f_n(x) / \int_0^1 f_n(t) dt$$

Synthèse de séquences

$$\begin{cases} s_\lambda = \lambda s_0 + (1 - \lambda) s_1 \\ \lambda \sim \mathcal{P}(g_n) \end{cases}$$



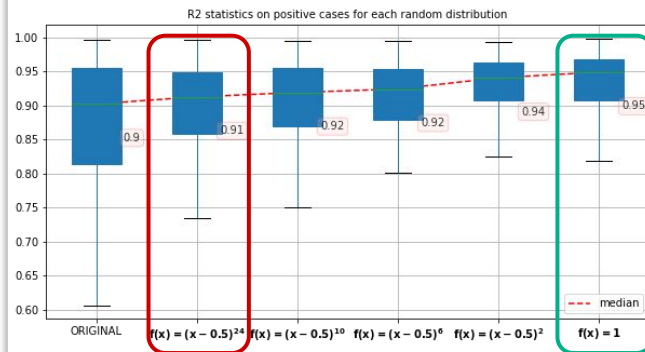
Données
d'entraînement

Sur-échantillonnage via g_n

CNN-1D

Résultats sur données synthétiques

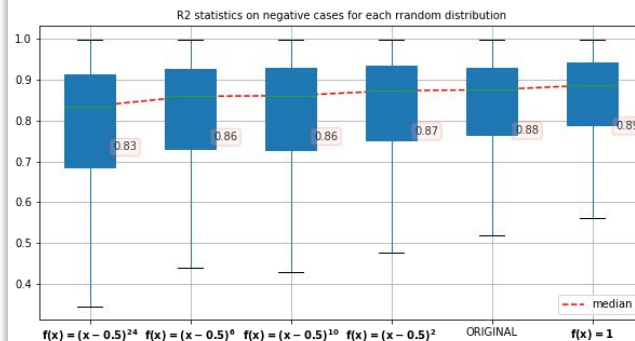
R2 sur cas minoritaires



La distribution
uniforme donne les
meilleurs résultats

La réplication
de données
donne de moins
bons résultats

R2 sur cas majoritaires



Problématique

- Contrôle qualité
- Variable cible à prédire : taux d'impureté
- Évènements rares à anticiper : augmentations brutales du taux d'impureté

Caractéristiques du jeu de données

- 10 variables (cible incluse)
- 90 points en entrées, 20 points en sortie
- 74000 échantillons
- 77% pour l'entraînement, 23% pour le test
- Environ 7% de cas rares

- Amélioration des résultats sur cas rares
- Baisse de performance sur cas majoritaires

	Cas positifs		Cas négatifs	
	MAE	RMSE	MAE	RMSE
Original	0.36 ± 0.37	0.46 ± 0.43	0.16 ± 0.21	0.19 ± 0.23
SMOTEST	0.34 ± 0.35	0.42 ± 0.40	0.19 ± 0.21	0.22 ± 0.23

TAB. 1: Résultats sur les données industrielles de test ($w_{in} = 90, w_{out} = 20$). Les valeurs à gauche et à droite du signe \pm correspondent resp. à la moyenne et l'écart type de l'erreur.

	TP	TN	FP	FN	bAcc	Rappel	Précision	Spécificité
Orig	3.87	89.88	3.22	3.03	76.31	56.08	54.58	96.54
OS + Unif	4.61	88.94	4.15	2.29	81.17	66.81	52.62	95.54

TAB. 6: Métriques de confusion sur le jeu de données industrielles (test).

Conclusions

Mode de rééchantillonnage

- Le sur-échantillonnage fournit les meilleurs résultats
- Le sous-échantillonnage rend le modèle sous-performant

Distribution aléatoire pour synthèse de séquences

- Une distribution uniforme fournit des résultats satisfaisants
- Une distribution proche de la duplication a tendance à rendre le modèle sous-performant

Données industrielles

- Augmentation des performances sur les cas d'intérêt

Perspectives

- Comparer d'autres méthodes d'augmentation de données
- Expérimenter sur plus de données industrielles
- Expérimenter avec d'autres modèles prédictifs (e.g. réseaux récurrents)
- Expérimenter sur d'autres types d'événements rares

Contacts



www.fieldbox.ai



melion@fieldbox.ai



[fieldboxai/predict-rare-events-smotest](https://github.com/fieldboxai/predict-rare-events-smotest)



Annexes

Caractérisation des cas rares

Ensemble de cas $\mathcal{D} : (X, Y) \in \mathbb{R}^{w_{in} \times n} \times \mathbb{R}^{w_{out}}$

- w_{in} : taille des séquences d'entrée
- w_{out} : taille des séquences de sortie
- n : nombre de signaux d'entrée
- $X \in \mathbb{R}^{w_{in} \times n}$: signaux d'entrée d'un cas donné
- $Y \in \mathbb{R}^{w_{out}}$: signal de sortie associé à X

Caractérisation d'un cas $\Phi : \mathbb{R}^{w_{out}} \mapsto \mathbb{R}$

Exemple : augmentation maximale

$$\Phi \left((y(t+k))_{k \in 1, w_{out}} \right) = \max_{\substack{t \in 1, w_{out}-1 \\ \tau \in 1, w_{out}-t}} y(t+\tau) - y(t)$$

Ensemble des cas rares

$$D_P = \{(X, Y) \in \mathcal{D} \mid \Phi(Y) \geq \theta\}$$

Evaluation

Métriques de régression

$$\text{RMSE}(Y, \hat{Y}) = \sqrt{\frac{1}{w_{out}} \sum_{t=1}^{w_{out}} (Y[t] - \hat{Y}[t])^2}$$

$$\text{MAE}(Y, \hat{Y}) = \frac{1}{w_{out}} \sum_{t=1}^{w_{out}} |Y[t] - \hat{Y}[t]|$$

Métriques de confusion

$$\text{Rappel}(Y, \hat{Y}) = \frac{TP}{TP + FN}$$

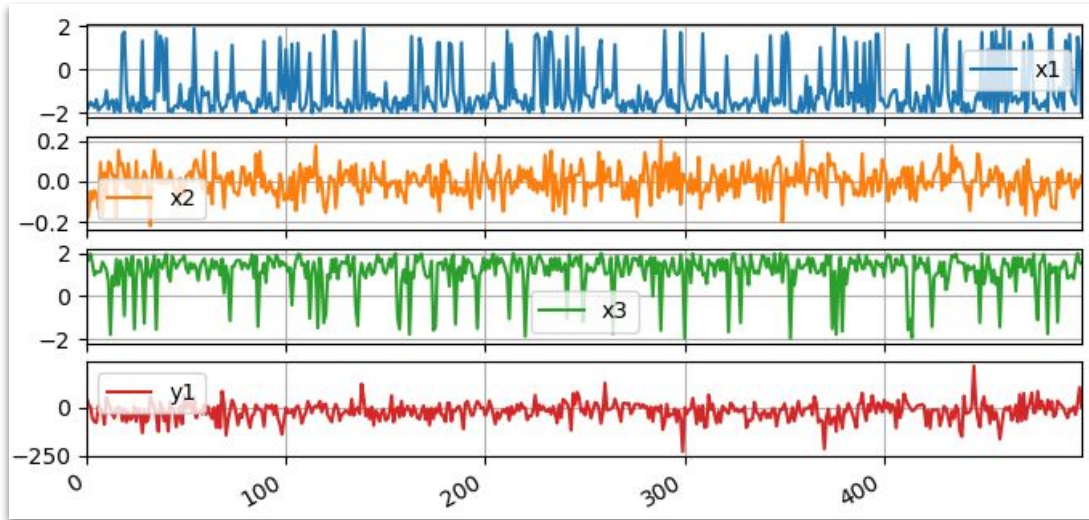
$$\text{Précision}(Y, \hat{Y}) = \frac{TP}{TP + FP}$$

$$\text{Spécificité}(Y, \hat{Y}) = \frac{TN}{TN + FP}$$

$$\text{bAcc}(Y, \hat{Y}) = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right)$$

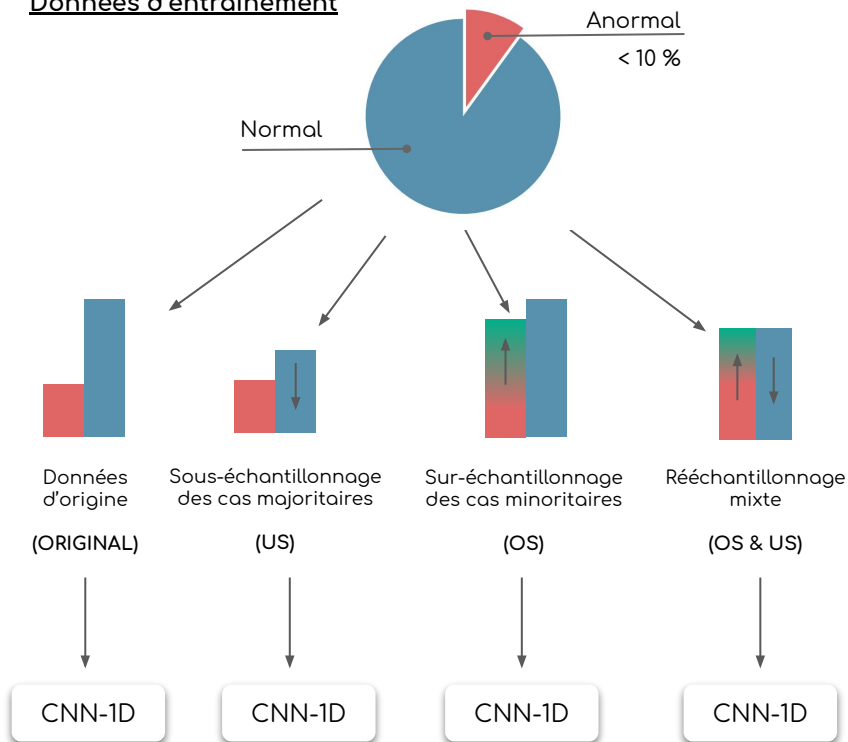
Caractéristiques du jeu de données

- 4 variables (cible incluse)
- 15 points en entrées, 5 points en sortie
- 9960 échantillons
- 70% pour l'entraînement, 30% pour le test
- Moins de 10% de cas rares

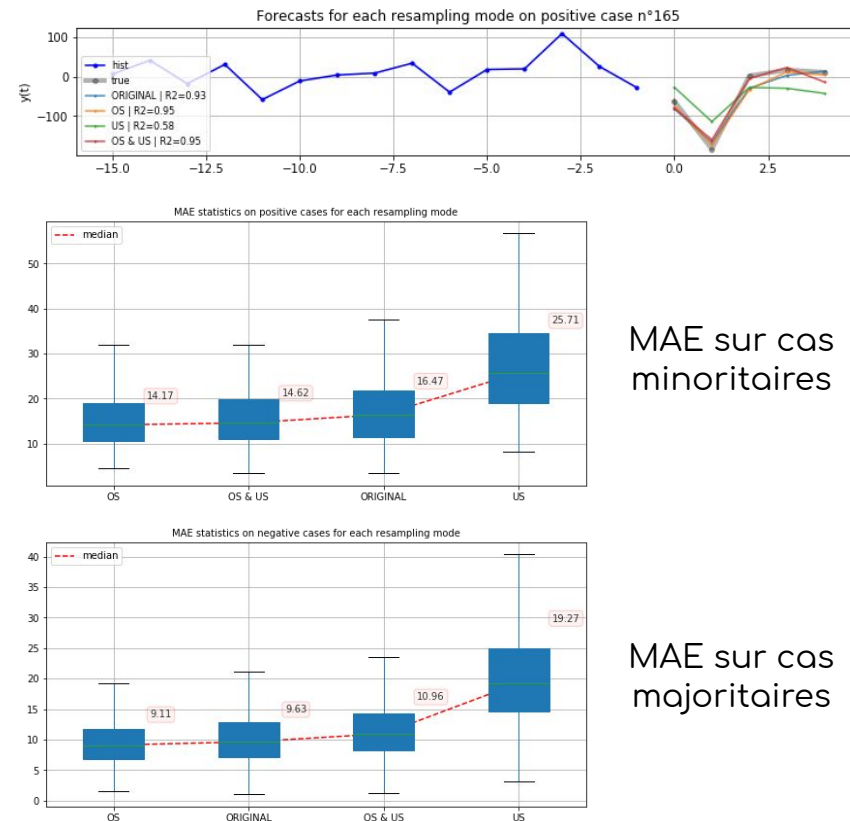


Influence du mode de rééchantillonnage

Données d'entraînement

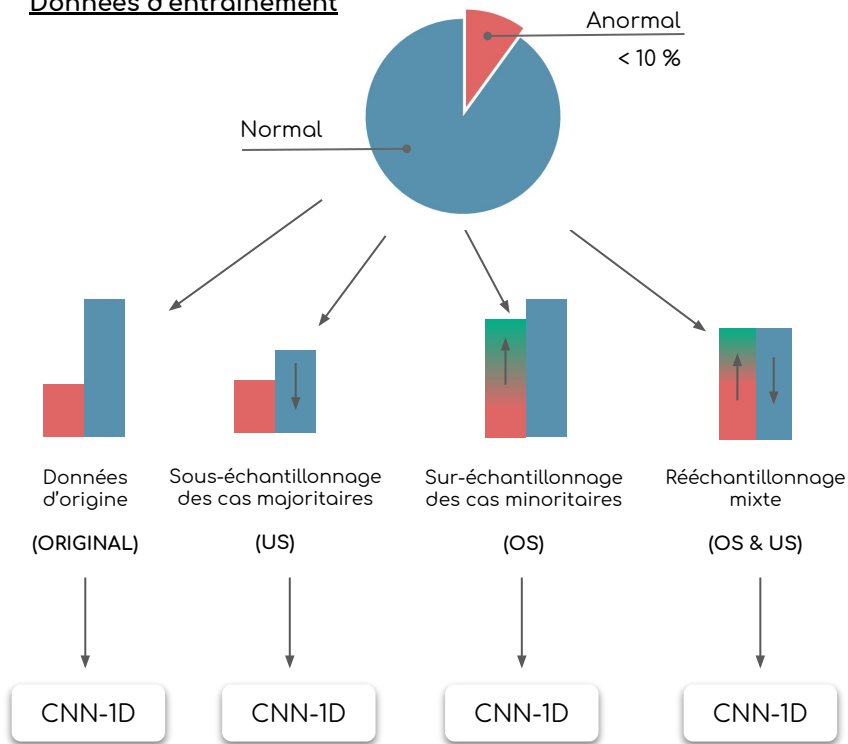


Résultats sur données synthétiques

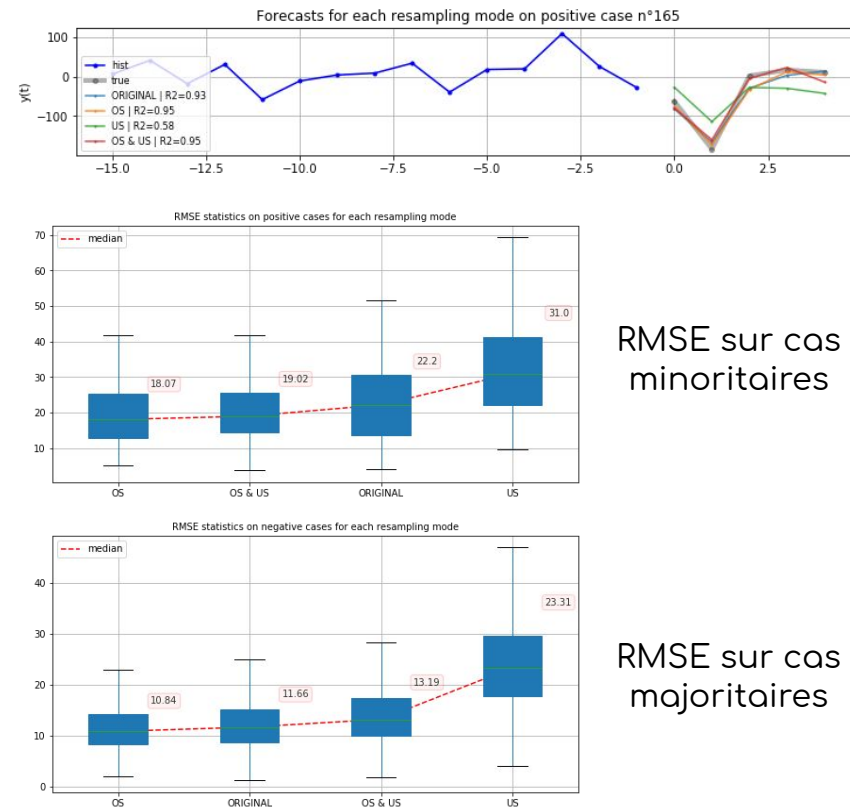


Influence du mode de rééchantillonnage

Données d'entraînement

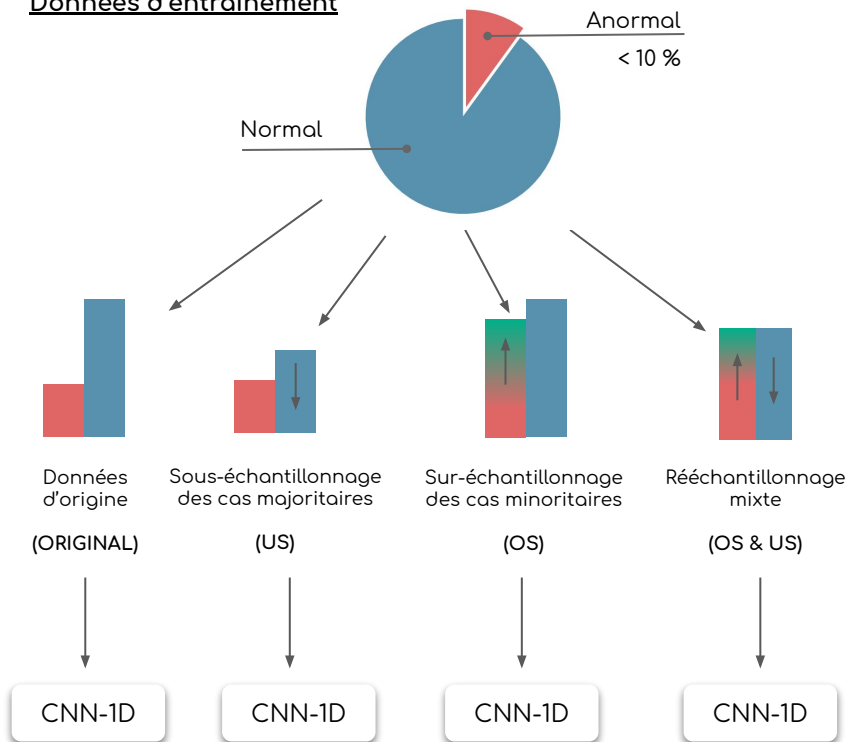


Résultats sur données synthétiques

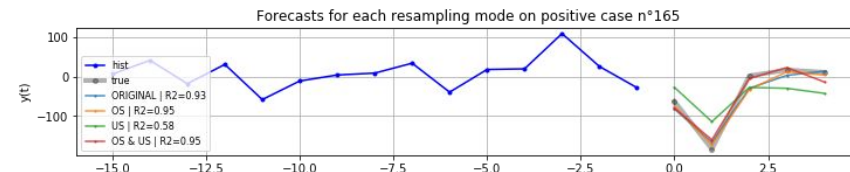


Influence du mode de rééchantillonnage

Données d'entraînement



Résultats sur données synthétiques



Métriques de confusion sur les données de test

	P	N	TP	TN	FP	FN	accuracy	recall	precision	specificity
ORIGINAL	8.43	91.57	4.15	90.37	1.2	4.28	94.52	49.21	77.5	98.69
OS	8.43	91.57	5.22	89.97	1.61	3.21	95.18	61.9	76.47	98.25
US	8.43	91.57	3.88	85.69	5.89	4.55	89.57	46.03	39.73	93.57
OS & US	8.43	91.57	5.55	89.36	2.21	2.88	94.92	65.87	71.55	97.59

Métriques de confusion sur les données d'entraînement

	P	N	TP	TN	FP	FN	accuracy	recall	precision	specificity
ORIGINAL	7.86	92.14	4.35	91.32	0.82	3.51	95.67	55.29	84.17	99.11
OS	47.74	52.26	41.39	51.51	0.75	6.34	92.9	86.71	98.21	98.56
US	50	50	31.66	48.54	1.46	18.34	80.2	63.32	95.59	97.08
OS & US	48.03	51.97	42.2	51.01	0.96	5.83	93.21	87.86	97.78	98.15

Influence de la distribution aléatoire

Fonctions forme

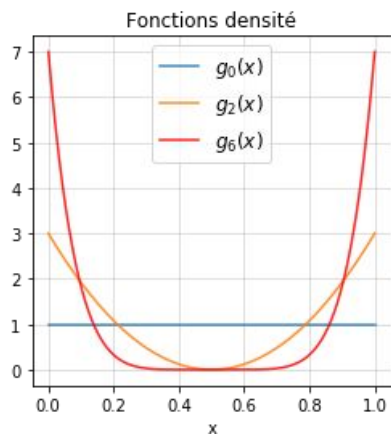
$$f_n(x) = (x - 0.5)^n$$

Fonctions densité

$$g_n(x) = f_n(x) / \int_0^1 f_n(t) dt$$

Synthèse de séquences

$$\begin{cases} s_\lambda = \lambda s_0 + (1 - \lambda) s_1 \\ \lambda \sim \mathcal{P}(g_n) \end{cases}$$



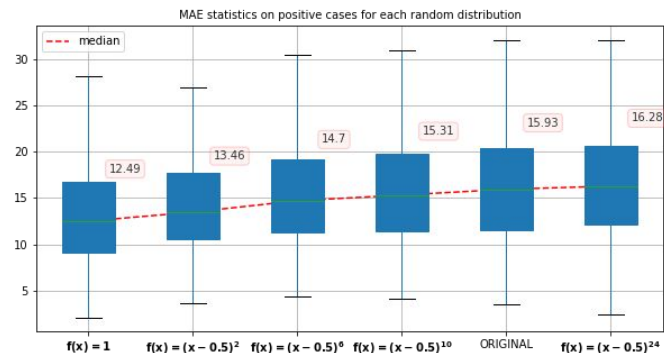
Données
d'entraînement

Sur-échantillonnage via g_n

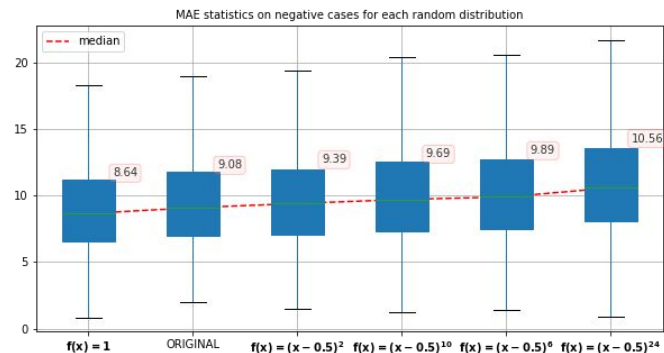
CNN-1D

Résultats sur données synthétiques

MAE sur cas minoritaires



MAE sur cas majoritaires



Influence de la distribution aléatoire

Fonctions forme

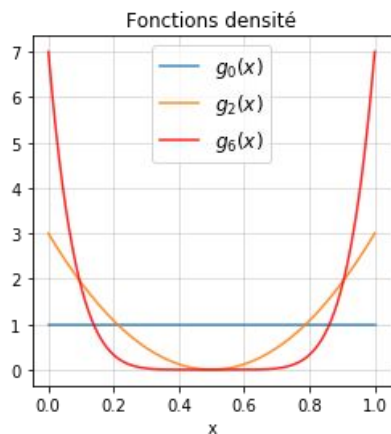
$$f_n(x) = (x - 0.5)^n$$

Fonctions densité

$$g_n(x) = f_n(x) / \int_0^1 f_n(t) dt$$

Synthèse de séquences

$$\begin{cases} s_\lambda = \lambda s_0 + (1 - \lambda) s_1 \\ \lambda \sim \mathcal{P}(g_n) \end{cases}$$



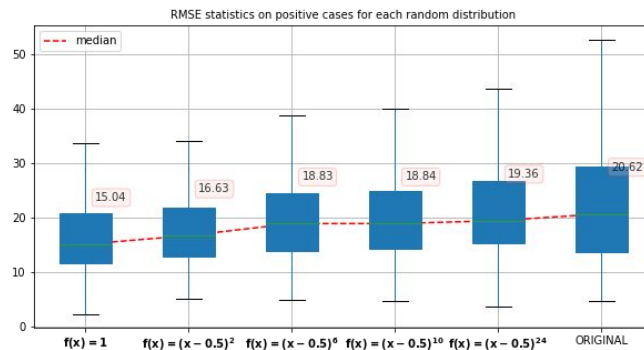
Données
d'entraînement

Sur-échantillonnage via g_n

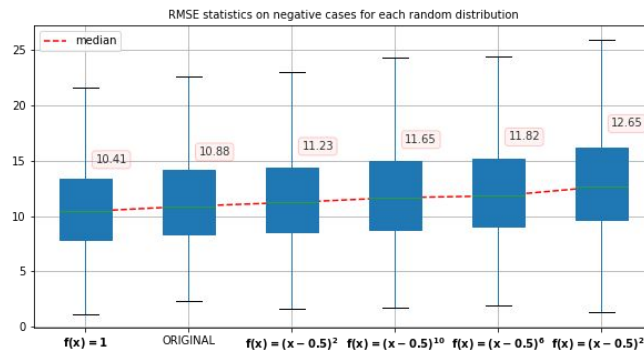
CNN-1D

Résultats sur données synthétiques

RMSE sur cas minoritaires



RMSE sur cas majoritaires



Influence de la distribution aléatoire

Fonctions forme

$$f_n(x) = (x - 0.5)^n$$

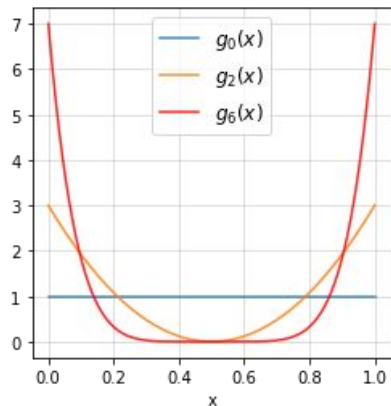
Fonctions densité

$$g_n(x) = f_n(x) / \int_0^1 f_n(t) dt$$

Synthèse de séquences

$$\begin{cases} s_\lambda = \lambda s_0 + (1 - \lambda) s_1 \\ \lambda \sim \mathcal{P}(g_n) \end{cases}$$

Fonctions densité

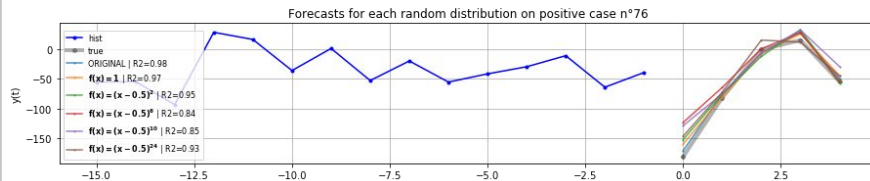


Données
d'entraînement

Sur-échantillonnage via g_n

CNN-1D

Résultats sur données synthétiques



Métriques de confusion sur les données de test

	P	N	TP	TN	FP	FN	accuracy	recall	precision	specificity
Original and Resampled Datasets										
ORIGINAL	8.43	91.57	3.75	90.43	1.14	4.68	94.18	44.44	76.71	98.76
$f(x) = 1$	8.43	91.57	5.55	90.5	1.07	2.88	96.05	65.87	83.84	98.83
$f(x) = (x - 0.5)^2$	8.43	91.57	5.42	90.17	1.4	3.01	95.59	64.29	79.41	98.47
$f(x) = (x - 0.5)^6$	8.43	91.57	5.82	89.77	1.81	2.61	95.59	69.05	76.32	98.03
$f(x) = (x - 0.5)^{10}$	8.43	91.57	4.75	90.7	0.87	3.68	95.45	56.35	84.52	99.05
$f(x) = (x - 0.5)^{24}$	8.43	91.57	4.62	90.77	0.8	3.81	95.38	54.76	85.19	99.12

Métriques de confusion sur les données d'entraînement

	P	N	TP	TN	FP	FN	accuracy	recall	precision	specificity
Original and Resampled Datasets										
ORIGINAL	7.86	92.14	4.65	91.05	1.09	3.21	95.7	59.12	81	98.82
$f(x) = 1$	47.94	52.06	41.02	51.42	0.63	6.92	92.44	85.56	98.48	98.78
$f(x) = (x - 0.5)^2$	47.9	52.1	41.38	51.5	0.59	6.52	92.88	86.39	98.58	98.86
$f(x) = (x - 0.5)^6$	48.06	51.94	42.68	50.86	1.08	5.38	93.54	88.81	97.54	97.93
$f(x) = (x - 0.5)^{10}$	47.99	52.01	39.42	51.3	0.71	8.57	90.72	82.15	98.24	98.64
$f(x) = (x - 0.5)^{24}$	48.02	51.98	40.78	51.04	0.94	7.24	91.82	84.93	97.75	98.19