Prévisions géographiques du nombre d'interventions des pompiers respectant la confidentialité différentielle locale

H. H. Arcolezi[1], J-F. Couchot[1], S. Cerna[1], C. Guyeux[1], G. Royer[2], B. Al Bouna[3], X. Xiao[4]

[1] Femto-ST Institute, Univ. Bourgogne Franche-Comté, France
[2] SDIS 25 - Service Dépar. d'Incend. et de Secours du Doubs
[3] TICKET Lab., Antonine University, Hadat-Baabda, Lebanon
[4] School of Computing, National University of Singapore, Singapore

APIA conférence, 2 juillet 2020

Plan

Plan

Motivations

Transport médical d'urgence en crise

- 80% de l'activité des Sapeurs Pompiers.
- En augmentation de 100% en moins de 10 ans à effectifs $\approx$ constants.
- Optimiser l'utilisation des ressources ? Renforcer (resp. réduire) les effectifs en périodes de pointe (resp. creuses).
- Nécessité d'avoir une vision des besoins à court, moyen et long terme $\rightsquigarrow$ apprentissage sur des données.

Des données publiques anonymisées trop fortement/faiblement [1]

- Trop fort : agrégation mensuelle ou hebdomadaire $\rightsquigarrow$ 12/52 points par an. Utilité des données ?
- Trop faible : fuite d'information pour les minuscules communes avec très peu d'intervention.

Objectif : résoudre le dilemme

- Garantir l'anonymat sur les données.
- Prédire correctement les interventions par apprentissage automatique.

1. https://www.data.gouv.fr/fr/datasets/interventions-des-pompiers-od71/ et https://www.data.gouv.fr/fr/datasets/interventions-des-pompiers/

Les données

Présentation des données du SDIS 25

ID	SDate	Caserne	Ville	Lieu
8	2008/08/08 08 :08	Besançon Est	Besançon	(47.2380, 6.0243)

- *ID* : identifiant d'intervention.
- *SDate* : date de début de l'intervention.
- *Caserne* : nom de la caserne qui a participé à l'intervention.
- *Ville* : nom de la municipalité où l'opération a eu lieu.
- *Lieu* : le lieu précis (latitude, longitude) de l'intervention.

Premières analyses

- Des données très critiques.
- 382046 interventions de pompiers entre 2006 à 2018 dans le Doubs.
- Nombre d'interventions : 17333 (2006) $\rightarrow$ 40510 (2018).

Confidentialité différentielle locale

Confidentialité différentielle (CD)

- $\mathcal{A}$: algorithme publiant des analyses issues de données privée.
- Intuition : $\mathcal{A}$ est CD ssi de n'importe quelle analyse on ne peut pas déduire que les données d'un individu particulier ont été utilisées dans celle-ci.
- Inconvénient : anonymisation réalisée lors de l'analyse $\rightsquigarrow$ des données immuables et stockées de manière sure.

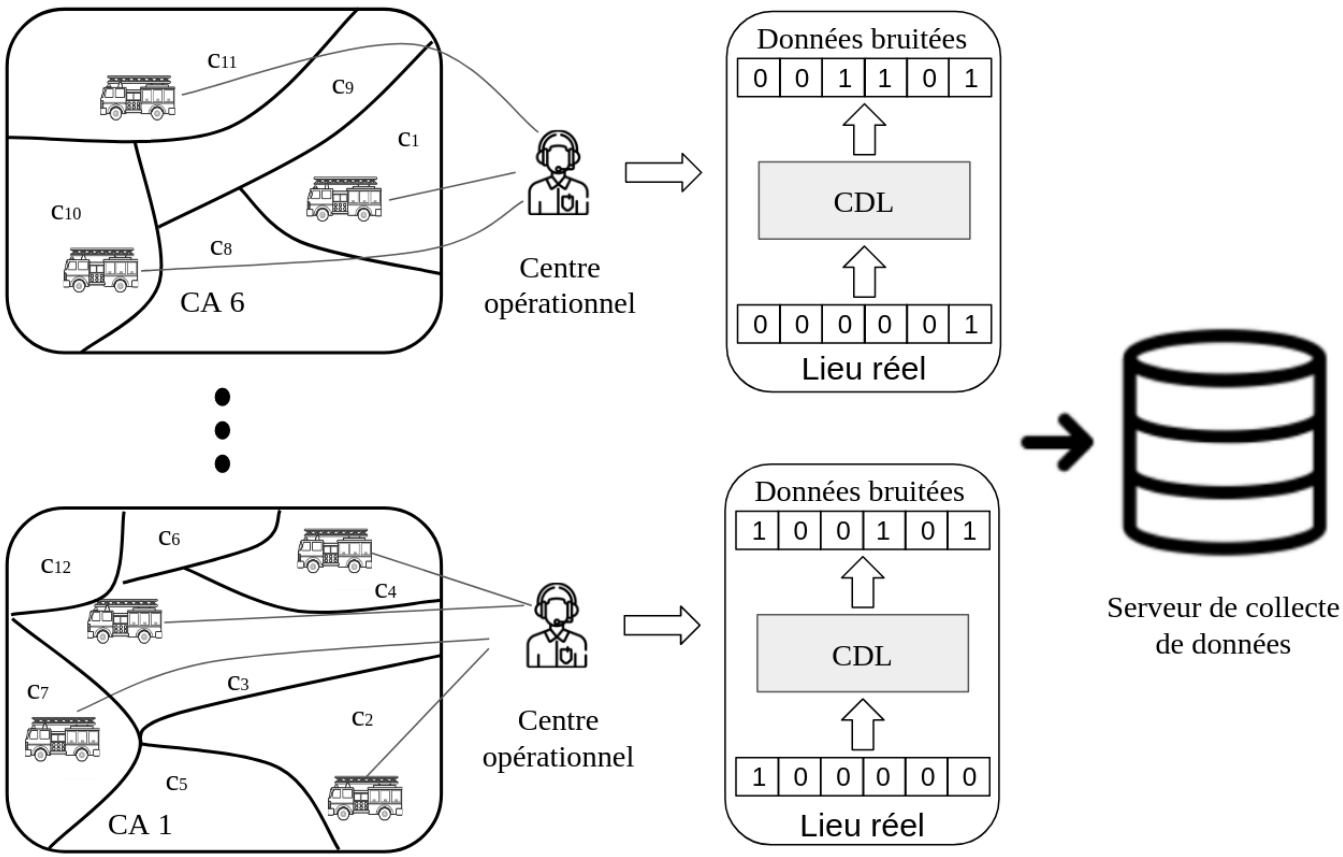
Confidentialité différentielle Locale (CDL)

- Données aseptisées par l'utilisateur de manière probabiliste avant d'être envoyées au collecteur.
- $\Pr[\mathcal{A}(v_1) \in R] \leq e^{\epsilon} \times \Pr[\mathcal{A}(v_2) \in R]$
  pour toutes données $v1$ et $v2$ et tous sous-ensembles R de $\mathrm{im}\mathcal{A}$ avec :
  - ϵ : nombre positif indiquant le niveau de fuite.
  - $\Pr[\mathcal{A}(v_1) \in R]$ la probabilité que l'analyse issue de la donnée v_1 soit un élément de R selon $\mathcal{A}$.

Plan

Schéma général

Après agrégation par Communauté d'Agglomération (CA).



RAPPOR [2] pour garantir le respect de la VP

Instanciation du RAPPOR basique

- **Donnée brute :** vecteur $B = (0, \ldots, 0, 1, 0 \ldots, 0)$ où $B_k = 1$ si le lieu de l'intervention est la $k^{\text{ème}}$ CA.
- **Réponse aléatoire permanente U transmise :** perturbation de chaque bit dans B selon

$$U_k = \begin{cases} 1 & \text{avec probabilité } \frac{1}{2}f, \\ 0 & \text{avec probabilité } \frac{1}{2}f \\ B_k & \text{avec probabilité } 1-f, \end{cases} \quad \text{et} \quad \text{avec } f \in [0,1]. \tag{1}$$

- **Niveau de fuite :** au pire égal à ϵ_∞

$$\epsilon_\infty = 2 \ln \left(\frac{1 - \frac{1}{2}f}{\frac{1}{2}f} \right). \tag{2}$$

2. Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor : Randomized aggregatable privacy-preserving ordinal response.In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 1054–1067, New York, NY, USA, 2014. ACM.

Estimation du nombre d'interventions par lieu

Approche statistique

- Nombre d'interventions (sur)estimé pour la localisation k, $NBint_{est}(k)$:

$$NBint_{est}(k) = \max\left(0; \frac{1}{1-f} \cdot \left(N_k - \frac{f \cdot N_{total}}{2}\right)\right) \tag{3}$$

- N_k : nbre de fois que le $k^{\text{ème}}$ bit vaut 1.
- N_{total} : nombre total d'interventions.

Évaluation de l'erreur

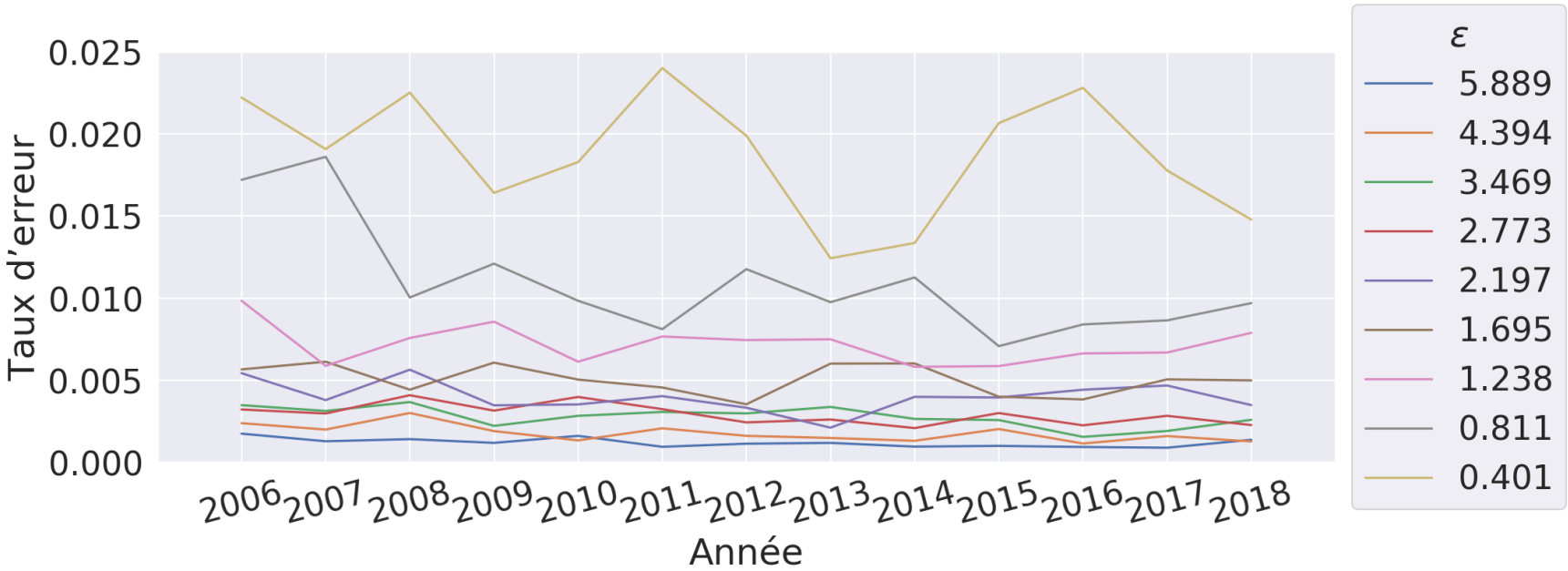
- Estimation de la densité de la $k^{\text{ème}}$ CA :

$$Density_{est}(k) = \frac{NBint_{est}(k)}{\sum_{j=1}^{17} NBint_{est}(j)} \tag{4}$$

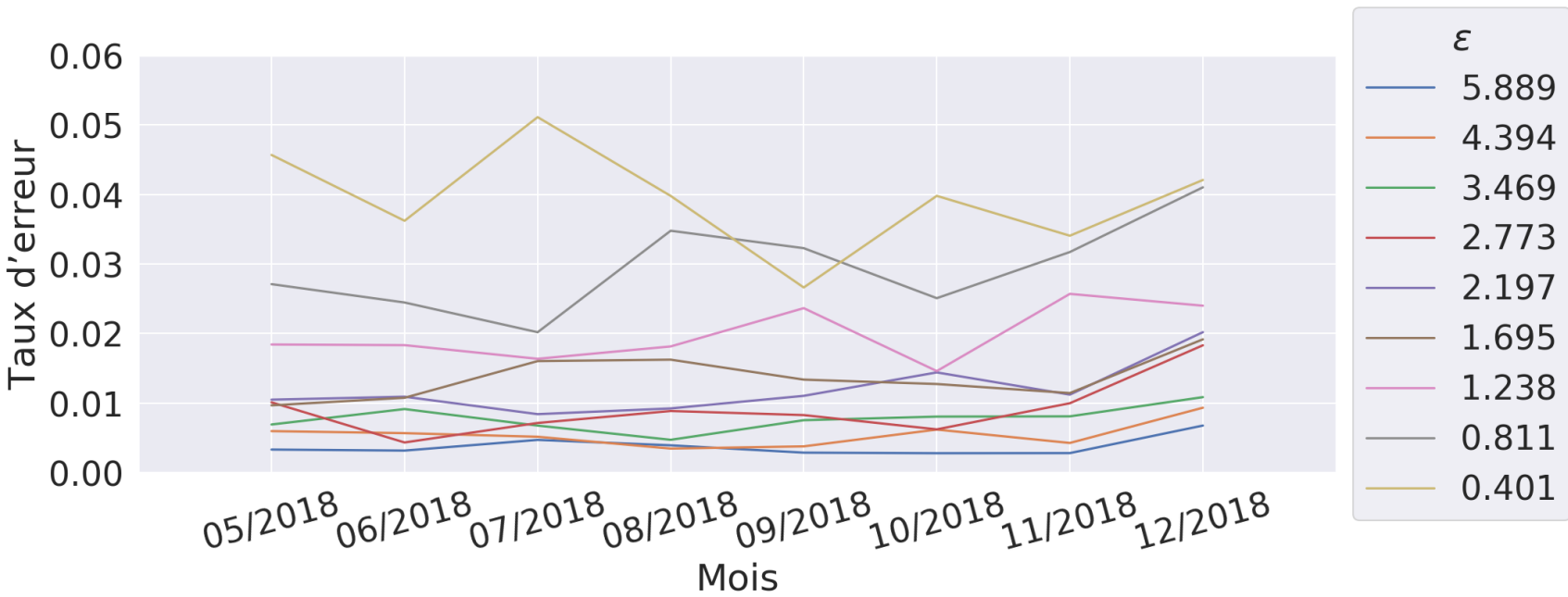
- Taux d'erreur défini par :

$$ER = \frac{1}{17} \sum_{j=1}^{17} |Density_{actuelle}(j) - Density_{est}(j)| \tag{5}$$

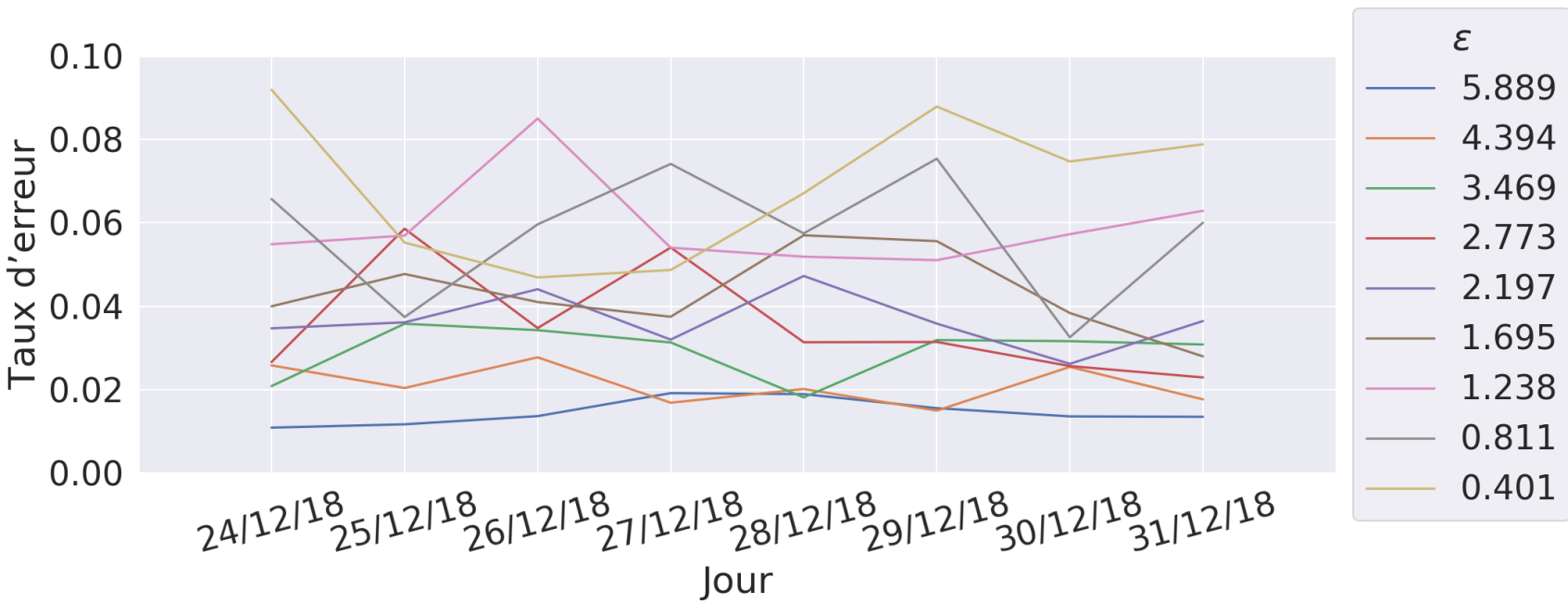
Evaluation du respect de la VP : taux d'erreur



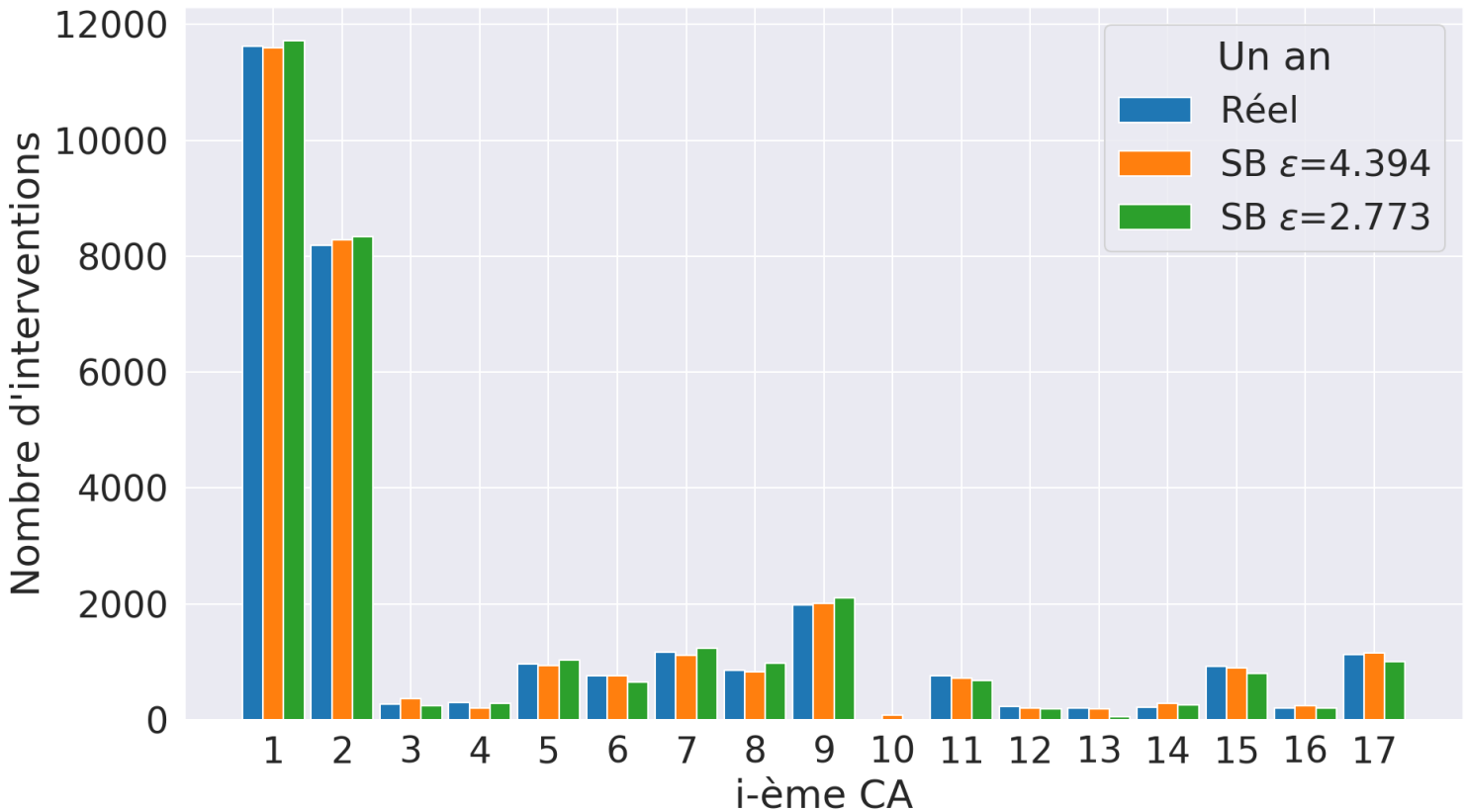
Evaluation du respect de la VP : taux d'erreur



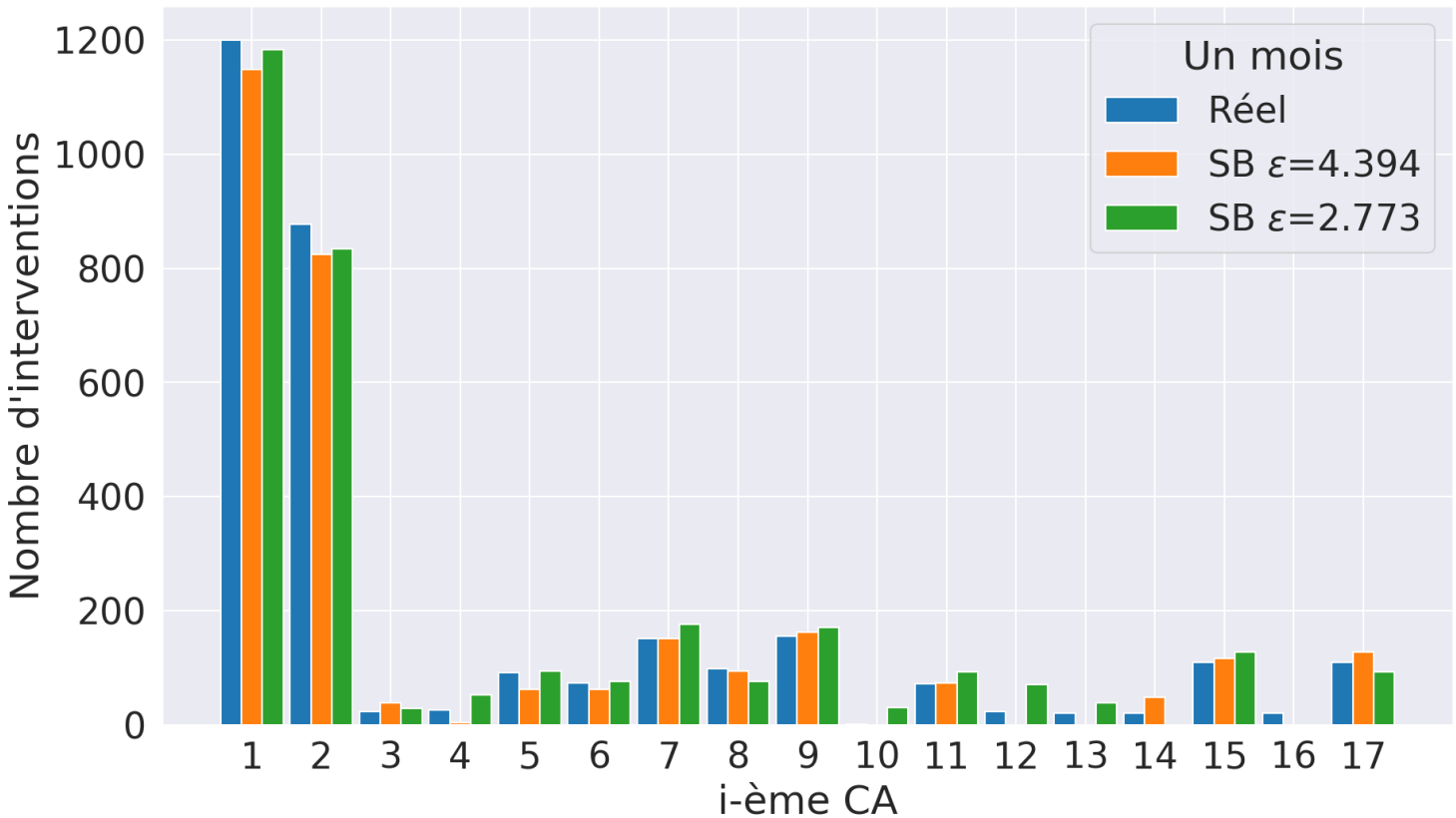
Evaluation du respect de la VP : taux d'erreur



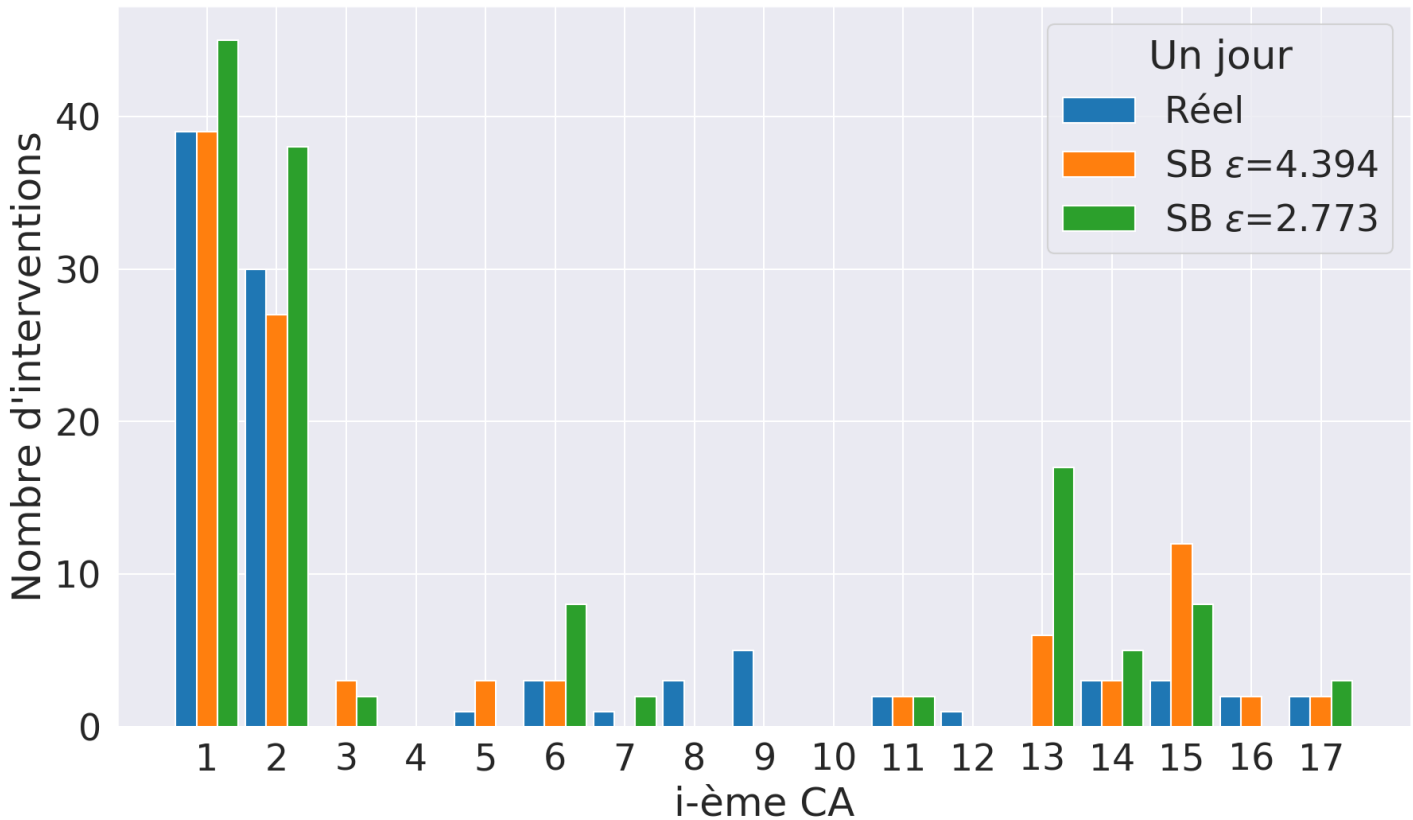
Evaluation du respect de la VP : estimations



Evaluation du respect de la VP : estimations



Evaluation du respect de la VP : estimations



Plan

Modèle d'apprentissage

Préparation des données

- Descriptions géométriques (polygones) des 17 CA du Doubs.
- Interventions des SP :
  - Données réelles (pour comparaison).
  - Données anonymes avec des f variant dans $[0,2; 0,4; 0,6; 0,8]$ (*i.e.*, ϵ_∞ dans $[4,394; 2,773; 1,695; 0,811]$).
- Informations temporelles : année, mois, jour, jour de la semaine, jour de l'année, premier/dernier jour du mois, de l'année.

Redressement des prévisions

- Nombres d'interventions surestimés par la méthode d'anonymisation.
- Sur toutes les données du SDIS calcul du ratio r^t de l'année précédente t :

$$r^t = \frac{\sum_{j=1}^{17} NBint_{est}^t(j)}{N_{total}^t} \tag{6}$$

- Pour chaque CA j, redressement de $NBint_{est}^{t+1}(j)$ en divisant par r^t.

Méthodologie d'apprentissage

- Régression multiple (MultiOutputRegressor de scikit-learn [3]).
- Un régresseur XGBoost par CA uniquement paramétré avec `objective = 'count:poisson`.
- Apprentissage sur 2006,. . . , 2017 et évaluation sur 2018.

3. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

Prédictions : erreurs et redressement

Comparaison de l'erreur quadratique moyenne (EQM) et de l'erreur absolue moyenne (EAM).

Modèle	Avec redressement		Ss. redressement	
	EAM	EQM	EAM	EQM
Modèle de base (moyenne)	-	-	2,5556	3,3237
Original	-	-	1,8552	2,5821
$f = 0,20$ $\epsilon_\infty = 4,39$	1,8666	2,5963	2,1748	2,8822
$f = 0,40$ $\epsilon_\infty = 2,77$	1,9271	2,7194	2,7436	3,6736
$f = 0,60$ $\epsilon_\infty = 1,69$	**1,9151**	**2,6848**	4,2475	4,9567
$f = 0,80$ $\epsilon_\infty = 0,81$	1,9403	2,7002	7,8542	8,4985

TABLE – Erreurs de prédiction du nombre d'interventions journalières par CA en 2018 en utilisant des données originales et anonymisées avec ou sans redressement.

Prédictions : un exemple journalier

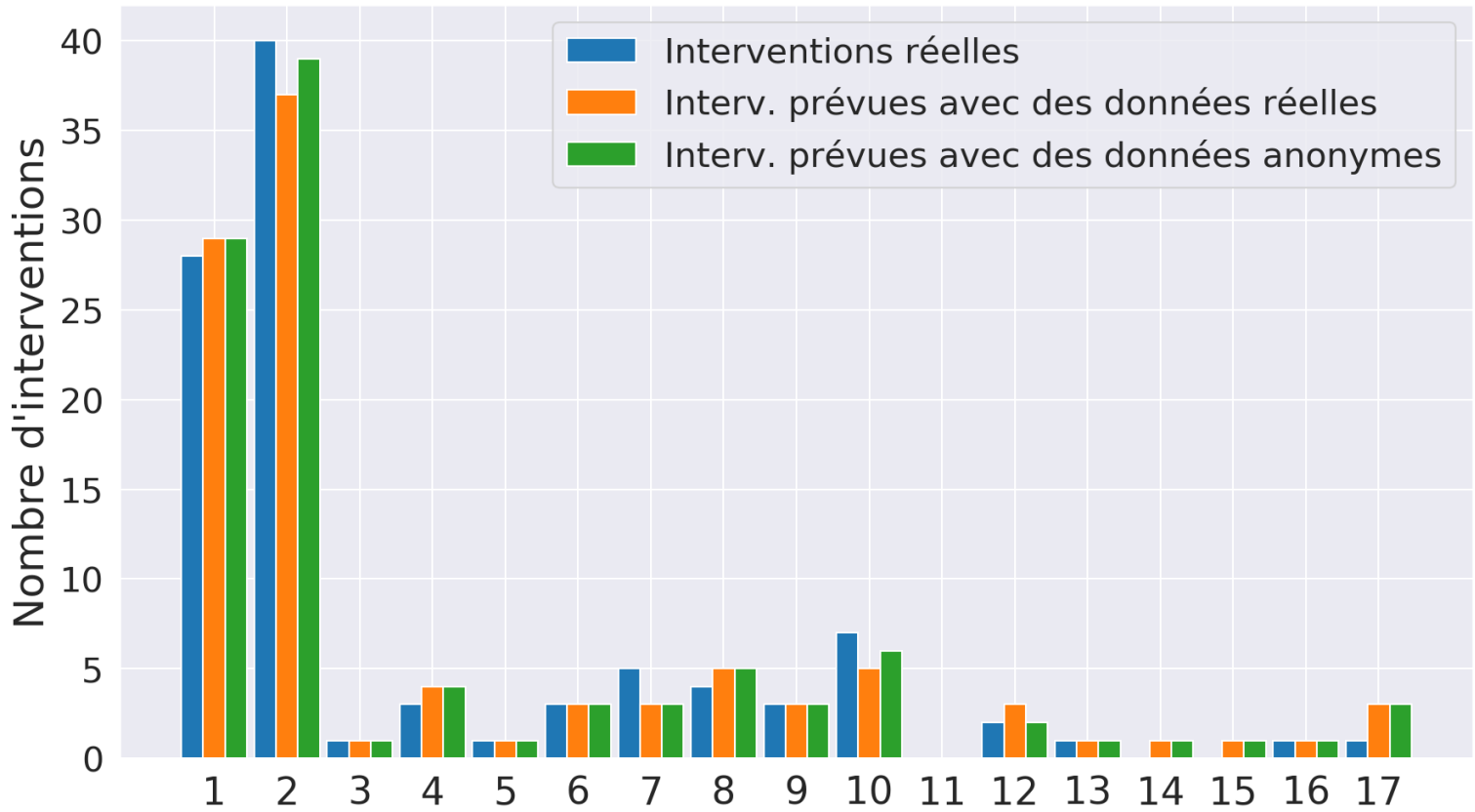


FIGURE – Comparaison entre les nombres réels et prévus d'interventions par CA pour un journée de mars 2018 avec $\epsilon_\infty = 1,69$ ($f = 0,60$)

Plan

Conclusion et perspectives

Synthèse

- Application d'un mécanisme de CDL pour la collecte respectueuse de données compatible avec
- Des prévisions précises par XGboost du nombre d'interventions par CA.
- Un article complet à lire [4]

Perspective

- Influence de l'étape d'agrégation des données à étudier.
- Augmentation fictive du volume des données pour l'apprentissage ?

4. Arcolezi, H. H., Couchot, J. F., Cerna, S., Guyeux, C., Royer, G., Al Bouna, B., & Xiao, X. (2020). Forecasting the Number of Firefighters Interventions per Region with Local-Differential-Privacy-Based Data. Computers & Security, 101888