Prévisions geographiques 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, UBFC, CNRS, Belfort France
[2] SDIS 25 - Service Départemental d'Incendie et de Secours du Doubs, France
[3] Lab., Antonine University, Hadath-Baabda, Lebanon
[4] School of Computing, National University of Singapore, Singapore

Résumé

Les travaux présentés ici visent à prédire le nombre d'interventions des pompiers par Communauté d'Agglomération tout en respectant la vie privée des utilisateurs. Une approche basée sur la confidentialité différentielle locale a été développée pour anonymiser les données de localisation, puis une approche d'apprentissage supervisé a été mise en place pour faire les prédictions. Les expériences réalisées montrent que la méthode complète d'anonymisation-prédiction est très précise : l'introduction de bruit n'affecte pas la qualité des prédictions, et ces dernières reflètent avec une bonne précision ce qui s'est passé dans la réalité.

Mots-clés

confidentialité différentielle locale, RAPPOR, lieu d'intervention des pompiers, prévision multi-cibles, XGBoost.

Abstract

The work presented here aims to predict the number of firefighters' interventions by region while respecting the privacy of users. An approach based on local differential privacy was used to anonymize the location data, then a supervised learning approach was used to make the predictions. The experiments carried out show that the complete anonymization-prediction method is very precise : the introduction of noise does not affect the quality of the predictions, and the predictions reflect with good accuracy what happened in reality.

Keywords

local differential privacy, RAPPOR, firemen intervention location, multi-target forecasting, XGBoost.

1 Introduction

Le transport médical d'urgence comprend les différents services utiles au transport des personnes blessées depuis leur domicile ou depuis le lieu de l'incident jusqu'à la structure médicale la plus apte à prendre soin du patient. La manière dont ce transport médical d'urgence est organisé dépend du pays considéré, de son histoire et des choix politiques qui ont été faits dans le passé. En France, par exemple, les Sapeurs Pompiers ne sont pas seulement responsables de l'extinction des incendies, mais ils doivent aussi prendre en charge une partie des transports médicaux d'urgence, et cette charge représente plus de 80% de leur activité.

Cette structuration a bien fonctionné dans le passé. Cependant, depuis un certain temps, en France comme dans divers autres pays, nous sommes confrontés à une crise majeure du transport médical d'urgence, pour diverses raisons (par exemple, le vieillissement de la population, la crise financière). Ces éléments et d'autres encore conduisent donc à une crise du transport sanitaire d'urgence dans diverses régions du monde.

Une des solutions envisagées pour soulager la pression sur ces transporteurs est d'optimiser l'utilisation de leurs ressources, afin de renforcer les équipes pendant les périodes de pointe, tout en les réduisant pendant les périodes creuses. Mais la situation de crise est telle qu'il est maintenant nécessaire d'aller beaucoup plus loin dans ces optimisations, ce qui nécessite une vision relativement claire des besoins à court, moyen et long terme.

Ainsi certains auteurs ont récemment cherché à exploiter les techniques d'intelligence artificielle [3, 2, 1, 9] afin de prévoir la demande future en matière de transport médical d'urgence. Cependant, pour être supervisé, l'apprentissage automatique nécessite d'avoir accès au flux d'intervention des opérateurs dont nous essayons de prédire la charge (ambulanciers privés, pompiers, etc.). Ces derniers n'ont généralement ni les ressources humaines et matérielles ni la compétence pour déployer des solutions basées sur l'intelligence artificielle et sont donc obligés de transmettre ces données à un tiers de confiance ayant cette compétence.

S'agissant de données sensibles liées à des accidents, au sauvetage de personnes, à des décès éventuels,..., comment ces données doivent-elles être publiées après avoir été anonymisées ? En France par exemple, deux bases de données récentes de tels flux ont été publiées sur `data.gouv.fr`. La première concerne les interventions 2007-2017 du Service Départemental d'Incendies et de Secours de Saône-et-Loire (SDIS 71) [1], contenant le nombre d'in-

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

terventions par type et par commune, tandis que la seconde concerne les mêmes types de données [2] du SDIS 91 (département de l'Essonne) pour la période 2010-2018. Dans chaque cas, l'anonymisation a été effectuée par agrégation : mensuelle pour la première série de données, et hebdomadaire pour la seconde.

Cependant, la manière dont ces données ont été diffusées pose deux problèmes : l'anonymisation obtenue est à la fois trop forte et trop faible. Trop forte, tout d'abord, parce que la réalisation d'une agrégation par mois entraîne une perte complète de l'utilité de celle-ci : elle résume les interventions à 12 points par an, pour lesquels seule une simple régression linéaire reste possible. Ensuite, trop faible, car cette agrégation par mois, ou par semaine, a été faite de manière aveugle et généralisée. Par exemple, dans le cas de l'agrégation mensuelle, il existe plus de 600 situations où il n'y a eu qu'une seule intervention dans une commune au cours d'un mois donné : à ce niveau, le simple 2-anonymat [13] n'est plus satisfait, et la fuite d'informations est évidente. Ces fuites d'informations sont également nombreuses dans le cas de données hebdomadaires, et l'anonymat a échoué pour les deux séries de données.

L'objectif de cet article est donc de montrer qu'il est possible de traiter de tels flux de manière à ce que l'anonymat soit garanti d'une part et que des prévisions correctes puissent être faites par apprentissage automatique sur ces données d'autre part. Ceci est vrai même si les données considérées ont des densités spatiales très variables. Plus spécifiquement, dans cet article, notre objectif est d'appliquer une version de la Confidentialité Différentielle Locale (CDL), afin de transformer les données réelles pour qu'elles soient à la fois correctement anonymisées, et utiles pour l'apprentissage automatique. Ces données traitées sont ensuite exploitées à des fins d'apprentissage et de prédiction : une tâche de prédiction du nombre d'interventions par communauté d'agglomération (CA), avec des données brutes et anonymes, est alors proposée. Les approches envisagées comprennent l'utilisation d'une mémoire longue à court terme pour le nombre total d'interventions [1] ; un perceptron multicouche pour le nombre total d'interventions à nouveau [9] ; et enfin l'utilisation de XGboost sur une tranche de temps de 3h, un modèle pour deux CA importantes, et des modèles par motif [2].

Le reste de cet article est organisé comme suit. La section suivante présente les données qui seront traitées dans cet article. La section 3 est consacrée au contexte théorique lié à la confidentialité différentielle et à sa version locale. Son application est expliquée à la section 4 et évaluée expérimentalement à la section 5. La capacité à effectuer des prédictions précises d'apprentissage machine sur la base de ces données est enfin détaillée dans la section 6. Cet article se termine par une section de conclusion, dans laquelle la contribution est résumée et des perspectives sont énoncées.

2. https://www.data.gouv.fr/fr/datasets/interventions-des-pompiers/

2 Présentation des données

La base de données à notre disposition a été fournie par le service d'incendie et de secours, SDIS 25, du département du Doubs (France). Ce fichier contient des informations sur 382046 interventions auxquelles les pompiers ont participé de 2006 à 2018 au sein de ce département. Chaque intervention est enregistrée dans un fichier sous la forme d'une ligne et les principaux attributs de ce fichier sont indiqués dans le tableau 1 avec des informations artificielles et décrits comme suit :

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

TABLE 1 – Principaux attributs des données relatives aux opérations des pompiers

- *ID* est l'identifiant d'intervention, qui est utilisé dans les dossiers supplémentaires ;
- *SDate* est la date de début de l'intervention ;
- *Caserne* est le nom de la caserne de pompiers qui a participé à l'intervention ;
- *Ville* est le nom de la municipalité où l'opération a eu lieu ;
- *Lieu* donne le lieu précis (latitude, longitude) de l'intervention.

En analysant les données, nous avons constaté une forte augmentation du nombre d'interventions au fil des ans. Autrement dit, en 10 ans, le nombre d'interventions a doublé de 17333 en 2006 à 34436 en 2016 et a continué d'augmenter jusqu'à 40510 en 2018.

3 Contexte théorique sur la confidentialité différentielle locale

Soit $\mathcal{A}$ un algorithme utilisé pour publier des informations issues d'une base de données privée. La confidentialité différentielle (CD) [6] est une contrainte sur $\mathcal{A}$ qui limite la divulgation d'informations privées des enregistrements se trouvant dans la base de données. Intuitivement, $\mathcal{A}$ est différentiellement privé si un observateur qui en voit un extrait ne peut pas dire si les informations d'un individu particulier ont été utilisées dans le calcul.
Soit ϵ un nombre réel positif qui correspond intuitivement au niveau de fuite. Plus la valeur de cette variable est élevée, plus la fuite d'informations est importante. Soit im $(\mathcal{A})$ représente l'image de $\mathcal{A}$, *i.e.*, l'ensemble de tous les résultats possibles par $\mathcal{A}$. On dit que l'algorithme $\mathcal{A}$ fournit ϵ -confidentialité différentielle si, pour tous les ensembles de données D_1 et D_2 qui diffèrent sur les données d'une personne, et pour tous les sous-ensembles R de im $(\mathcal{A})$, nous avons

$$\Pr[\mathcal{A}(D_1) \in R] \leq e^{\epsilon} \times \Pr[\mathcal{A}(D_2) \in R], \qquad (1)$$

avec $\Pr[\mathcal{A}(D_2) \in R]$ la probabilité qu'un ensemble de données D_2 puisse être anonymisé en un élément de R selon $\mathcal{A}$ et ϵ le montant de la fuite. Cette équation donne une limite supérieure de la probabilité qu'un ensemble de

données D_1 puisse être anonymisé en un élément de R, ce qui constitue donc une fuite d'informations. Le lecteur intéressé pourra lire [5, 7].
Toutefois, cette approche exige que l'ensemble des données soit complet et stocké de manière sure. L'anonymisation ne se fait pas avant. La Confidentialité Différentielle Locale (CDL) [8] est une solution à ce problème. Dans cette approche, les données sont aseptisées par l'utilisateur de manière probabiliste avant d'être envoyées au collecteur. Un exemple simple consiste à demander à une personne de répondre à la question "Habitez-vous à Belfort ?", selon la procédure suivante :
Lancez une pièce de monnaie.

- Si c'est "pile", lancez à nouveau la pièce (en ignorant le résultat) et répondez honnêtement à la question.
- Si c'est "face", alors relancez la pièce et répondez "Oui" si face et "Non" sinon.

Soit t_y la proportion de réponses "Oui" vraies et c_y la proportion de réponses "Oui" observées. L'équation suivante donne une relation estimée entre ces deux variables

$$\frac{1}{2}t_y + \frac{1}{4} \approx c_y.$$

Plus le nombre d'expériences est élevé, plus la proportion de réponses "Oui" aléatoires sera proche de 1/4 et plus le nombre de fois où la vérité est dite sera proche, plus l'estimation sera précise. Dans ce cas, t_y peut être estimé par

$$t_y \approx 2.c_y - \frac{1}{2}.$$

L'algorithme $\mathcal{A}$ est censé fournir ϵ-CDL si, pour toutes les paires de données privées possibles de l'utilisateur v_1 et v_2 et tous les sous-ensembles R de im$\mathcal{A}$:

$$\Pr[\mathcal{A}(v_1) \in R] \leq e^{\epsilon} \times \Pr[\mathcal{A}(v_2) \in R]. \quad (2)$$

4 Collecte de données sur la localisation des interventions des pompiers dans le respect de la vie privée

La première question que l'on peut se poser est de savoir si une intervention est un attribut sensible. La réponse est oui, car les pompiers n'auraient pas été appelés si la situation n'avait pas été suffisamment grave. Prenons par exemple le scénario où une personne, qui habite dans un petit village, a contracté une maladie très particulière. Si l'on sait que, pendant cette période, une intervention a eu lieu dans cette ville, alors qu'elle est normalement rare, il y a une forte probabilité que les pompiers soient intervenus pour cette personne.
Par conséquent, le but de cette tâche est de mettre en œuvre un mécanisme de préservation de la vie privée pour la localisation de l'intervention des pompiers en utilisant le concept de confidentialité différentielle locale décrit précédemment. Ensuite, pour évaluer le compromis vie privée/utilité, le défi consiste à estimer approximativement le nombre d'interventions des pompiers sur les lieux analysés en utilisant les données anonymes.
Les deux sous-sections suivantes présentent l'approche de préservation de la vie privée basée sur la CDL appliquée à la collecte de données de localisations des interventions des pompiers et la méthode statistique pour estimer le nombre d'interventions par localisation.

4.1 Approche de la collecte de données basée sur la CDL

La figure 1 illustre un aperçu de l'approche et est résumée dans ce qui suit.

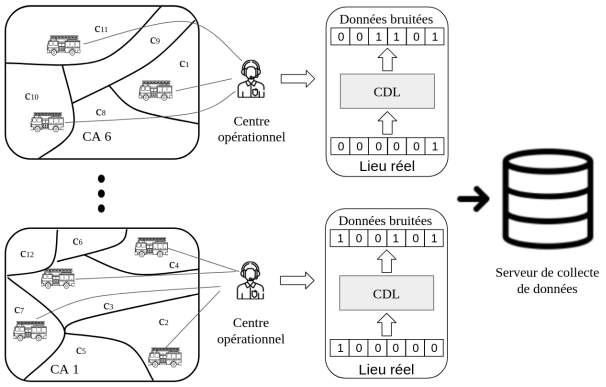

FIGURE 1 – Un aperçu de l'approche appliquée pour collecter les données de localisation des interventions des pompiers en préservant la vie privée.

Dans l'approche proposée, la première étape pour garantir la confidentialité du lieu de chaque intervention consiste à regrouper les villes qui ont fait l'objet de chaque intervention au niveau d'une communauté d'agglomération pour obtenir des événements suffisamment représentatifs en nombre. Par exemple, on peut remarquer dans la figure 1 qu'un ensemble de petites villes $C = \{c_1,c_2,...,c_{12},...,c_m\}$ sont regroupées en $n = 6$ CA.
Dans ce contexte, en utilisant les données dont nous disposons, 608 villes où des interventions ont eu lieu dans le département du Doubs sont agrégées en $n = 17$ CA en utilisant l'ensemble de données publiques disponibles [3]. Les 17 CA sont : (1) CA du Grand Besançon, (2) CA Pays de Montbéliard Agglomération, (3) CC Altitude 800, (4) CC de Montbenoit, (5) CC des Deux Vallées Vertes, (6) CC des Lacs et Montagnes du Haut-Doubs, (7) CC des Portes du Haut-Doubs, (8) CC du Doubs Baumois, (9) CC du Grand Pontarlier, (10) CC du Pays d'Héricourt, (11) CC du Pays de Maîche, (12) CC du Pays de Sancey-Belleherbe, (13) CC du Plateau de Frasne et du Val Rasne et du Val de Drugeon (CFD), (14) CC du Plateau de Russey, (15) CC du Val de Morteau, (16) CC du Val Marnaysien, (17) CC Loue-Lison.
Deuxièmement, pour améliorer le niveau de confidentialité de chaque intervention, le mécanisme "Basic One-time

3. https://www.collectivites-locales.gouv.fr/liste-et-composition-2018/

RAPPOR" introduit par [8] est appliqué. Cet algorithme est une simplification du mécanisme RAPPOR, qui utilise des filtres de Bloom et des fonctions de hachage pour cartographier les rapports envoyés par les utilisateurs et il comporte deux niveaux de réponses aléatoires, à savoir les réponses permanentes et les réponses instantanées.
Cependant, dans le "Basic One-time RAPPOR", il n'est appliqué qu'une seule étape de réponse aléatoire en utilisant une cartographie déterministe des 17 CA en tableau de bits. La motivation pour utiliser cet algorithme simple est la suivante : les communautés d'agglomération sont connues a priori en permettant la cartographie déterministe plutôt qu'en utilisant des fonctions de hachage et des filtres de Bloom ;
Une application technique de cet algorithme dans notre étude de cas est décrite ci-dessous :

1. **Vrai signal de localisation.** Soit $R = \{r_1,r_2,...,r_n\}$ un ensemble de n CA, où chaque indice représente un identifiant de CA unique. Ainsi, un tableau de n bits, B (qui indique le lieu d'intervention actuel) est défini comme

$$B_k = \begin{cases} 1 & \text{si } k = i \text{ et} \\ 0 & \text{sinon,} \end{cases} \tag{3}$$

   où B_k représente la valeur du $k^{\text{ème}}$ bit dans B, $k \in \{1,\ldots,n\}$. Autrement dit, le bit correspondant à l'identifiant des CA est mis à 1, tandis que les autres sont mis à 0.

2. **Réponse aléatoire permanente.** Ensuite, chaque bit dans B (de l'étape précédente) est perturbé en appliquant le concept de réponse aléatoire comme suit :

$$U_k = \begin{cases} 1 & \text{avec probabilité } \frac{1}{2}f, \\ 0 & \text{avec probabilité } \frac{1}{2}f \text{ et} \\ B_k & \text{avec probabilité } 1-f, \end{cases} \tag{4}$$

   où f est une valeur de probabilité entre 0 et 1, qui contrôle le niveau de garantie de confidentialité différentielle de ϵ.

3. **Rapport final.** La réponse aléatoire permanente B est transmise au serveur du collecteur de données.

Il a été démontré [8] que le niveau différentiel local de confidentialité ϵ est au pire ϵ_∞ défini comme

$$\epsilon_\infty = 2\ln\left(\frac{1-\frac{1}{2}f}{\frac{1}{2}f}\right). \tag{5}$$

4.2 Estimation du nombre d'interventions par lieu

En considérant un moment précis que l'on souhaite analyser, l'objectif est d'estimer approximativement le nombre d'interventions par lieu associé à la $i^{\text{ème}}$ CA, r_i. Soit donc $set(U)$ un ensemble de réponses permanentes randomisées et $set(B)$ l'ensemble correspondant de tableaux de bits de localisation d'origine. En outre, supposons que $|set(U)|$ et $|set(B)|$ indiquent le nombre d'éléments dans chaque ensemble respectif. Naturellement, $|set(U)| = |set(B)|$.
Par conséquent, le nombre estimé d'interventions $NBint_{est}$ par emplacement de CA r_i pour $i \in [1,n]$ est acquis par une approche basée sur les statistiques (SB) comme suit [8] :

$$NBint_{est}(r_i) = \frac{1}{1-f} \cdot \left(N_i - \frac{f \cdot N_{total}}{2}\right) \tag{6}$$

où N_{total} est le nombre de réponses permanentes randomisées $|set(U)|$ et N_i est le nombre total de réponses permanentes randomisées dont le $i^{\text{ème}}$ bit est fixé à 1. N_i est calculé en utilisant les données de $set(U)$. Il convient de noter que l'équation (6) peut estimer des nombres négatifs, d'où l'utilisation de la fonction $max(0,NBint_{est})$. Dans la littérature, d'autres techniques comme l'algorithme de maximisation des attentes [4] sont également utilisées pour estimer les densités.
Pour évaluer le résultat SB, l'estimation de la densité d'un emplacement de la $i^{\text{ème}}$ CA associée à r_i est calculée comme suit [11] :

$$Density_{est}(r_i) = \frac{NBint_{est}(r_i)}{\sum_{y=1}^{n} NBint_{est}(r_y)} \tag{7}$$

où n est le numéro de la CA, et, par conséquent, la métrique du taux d'erreur (*ER*) est définie comme

$$ER = \frac{1}{n}\sum_{i=1}^{n} |Density_{actual}(r_i) - Density_{est}(r_i)| \tag{8}$$

où $Density_{actual}(r_i)$ et $Density_{est}(r_i)$ correspondent respectivement à la densité réelle et à la densité estimée de la CA associée à la $i^{\text{ème}}$ localisation. Plutôt que de calculer la racine carrée de l'erreur quadratique moyenne sur le nombre estimé et réel d'interventions, le taux d'erreur est calculé sur la valeur de densité motivée par des valeurs normalisées comprises entre 0 et 1.

5 Expériences d'anonymisation

Pour évaluer l'approche consistant à anonymiser le lieu d'intervention des pompiers, plusieurs simulations sont réalisées avec différentes valeurs de f, ce qui détermine le niveau de ϵ_∞-différence de vie privée. Ainsi, en utilisant l'approche statistique (Equation (6)), l'objectif est d'estimer le nombre d'interventions par CA en considérant différents scénarios de temps. Ces expériences permettront d'évaluer la relation entre *ER* et la taille des données (période d'analyse) en fonction de f afin de trouver le meilleur compromis vie privée-utilité pour différentes applications.
Les scénarios de temps sont décrits ci-dessous. Chaque scénario permet aux pompiers de disposer d'une base de données anonyme où des entreprises tierces ou le département des ressources humaines lui-même pourraient acquérir des statistiques de grande utilité. Dans les expériences, f variera dans $[0{,}1; 0{,}2; \ldots; 0{,}8; 0{,}9]$, ce qui garantit ϵ_∞-une différence de confidentialité dans $[5{,}89; 4{,}39; \ldots; 0{,}81; 0{,}4]$ respectivement.

La première analyse co ncerne des données sur un an (13 points de données), ce qui permet au début d'une année aux pompiers de mieux répartir leur budget autour de leurs centres en fonction du nombre d'interventions par CA. Ensuite, un scénario d'un mois (156 points de données) est envisagé. Les pompiers peuvent ainsi disposer de statistiques pour réorganiser les budgets et le personnel chaque mois. Enfin, un scénario d'un jour (4748 points de données) est pris en considération de sorte que des tâches d'apprentissage machine puissent être appliquées dans cette quantité de données.

5.1 Resultats

Le but de cette tâche est d'anonymiser les données de localisation des interventions des pompiers avec différents niveaux de confidentialité (mesuré à l'aide de ϵ_∞) et trois scénarios de temps et d'évaluer les relations entre *ER* et la taille des données.

Pour mieux illustrer les résultats, la figure 2 montre la relation entre *ER* et ϵ_∞ pour chaque année (2006-2018), avec un zoom pour les 8 derniers mois de 2018, et avec un zoom pour les 8 derniers jours de décembre de 2018, respectivement. De plus, la figure 3 illustre les statistiques acquises sur le nombre d'interventions pour l'année 2013, le premier mois de 2017, et un jour précis du mois de janvier 2016 avec deux valeurs pour ϵ_∞ égal à 4,394 et 2,773 engendrées respectivement à partir de f égal à 0,2 et 0,4 en suivant l'équation (5).

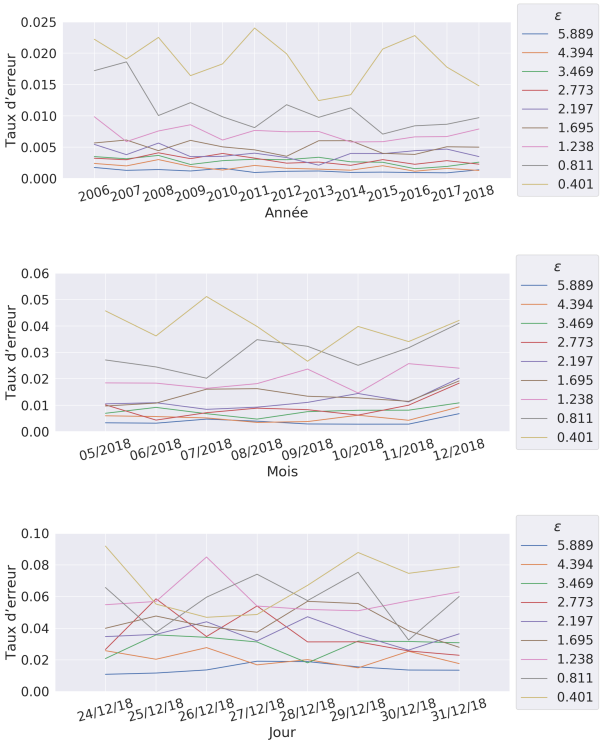

FIGURE 2 – Relation entre le taux d'erreur *ER* et la durée de l'étude variant ϵ_∞.

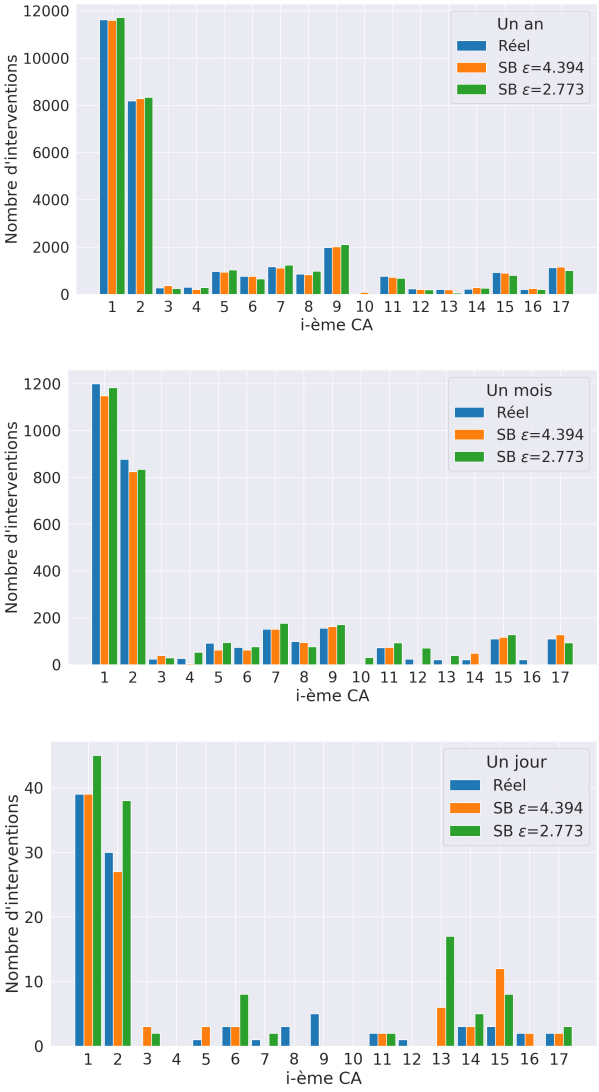

FIGURE 3 – Analyse entre le nombre réel et le nombre estimé d'interventions par CA.

5.2 Discussions

Comme on peut le remarquer dans les figures 2 et 3, le mécanisme basé sur la CDL peut être bien appliqué à la collecte de la localisation des interventions des pompiers dans le but de déduire le nombre d'interventions par CA. Comme la CDL garantit la confidentialité des personnes en perturbant les données avant de les envoyer au collecteur de données, dans ce cas, le pompier chargé de signaler les interventions appliquera la perturbation à la localisation des interventions avant de l'envoyer au serveur de données (comme l'illustre la figure 1).

Il est à noter que plus le nombre de points de données diminue, plus le *ER* augmente. Comme le souligne également la littérature, la CDL nécessite une grande quantité de données pour garantir un bon équilibre du bruit, car le bruit est ajouté à chaque point de données dans cette approche. Alors que pour une analyse sur un an, le nombre

d'interventions est d'au moins 17333 pour l'année 2006, la moyenne par jour n'est que de 47 pour la même année. Pour cette raison, l'utilité des données diminue si l'on veut des scénarios plus précis, comme les cas d'un mois et d'un jour présentés dans cet article.

De plus, la relation entre les *ER* et les garanties de confidentialité est naturelle : plus ϵ_∞ est petit, plus la garantie de confidentialité différentielle est renforcée, plus le bruit ajouté aux données est important et plus l'utilité de celles-ci est faible. Prenons par exemple la figure 3, particulièrement celle détaillant les prédictions journalières. On constate dans celle-ci que les prévisions issues d'une anonymisation avec ϵ_∞ petit (*i.e.* en vert sur la courbe) sont plus éloignées de la réalité que celles obtenues à la suite d'une anonymisation avec un ϵ_∞ plus grand.

Toutefois, comme nous l'avons déjà dit, la taille des données a une grande influence à ce stade.

Supposons que nous ayons une densité suivant une distribution normale $\textit{Density}_{actual}(r_i) = 1/n$. Un $ER = 0{,}01$ représente en moyenne 10% d'erreur par CA, r_i. La figure 2 montre la relation entre *ER* et la taille des données : pour le scénario d'un an (resp. un mois, une jour) avec le niveau maximum de garanties de confidentialité, le taux d'erreur *ER* oscille entre $0{,}001$ et $0{,}025$ (resp. entre $0{,}004$ et $0{,}05$, entre $0{,}01$ et $0{,}09$).

La Figure 3 montre les statistiques obtenues sur le nombre d'interventions pour chaque période : des petites erreurs sont obtenues pour le cas d'un an (ER normalement inférieur à $0{,}01$) ; des erreurs raisonnable pour le scénario d'un mois (la moitié de ER est inférieure à $0{,}01$), et des erreurs considérables pour le scénario d'un jour (tous $ER \geq 0{,}01$). Par conséquent, comme le souligne également la littérature, le choix de ϵ dépend de plusieurs facteurs (taille des données, domaine d'application) et il faut trouver un équilibre approprié entre confidentialité des données et utilité de celles-ci. Par exemple, dans ce cas, où les 608 villes ont été agrégées en $n = 17$ CA, certains jours avec peu de données (moins de ~ 100 par ex.), la confidentialité pourrait être légèrement diminuée afin d'en préserver l'utilité. Dans la littérature, les valeurs d'ϵ se situent classiquement dans la fourchette $[0{,}01; 10]$ [10].

6 Prévision des interventions des pompiers par CA

Le but de cette tâche est d'implémenter un algorithme d'apprentissage machine efficace, à savoir le renforcement du gradient extrême (XGBoost), pour prévoir le nombre d'interventions par jour des 17 CA du Doubs-France. L'objectif principal est d'utiliser des fichiers anonymes pour construire des modèles afin d'évaluer l'utilité des données avec différents niveaux de confidentialité ϵ_∞ différents par rapport à l'original.

6.1 Préparation des données

Trois sources initiales ont été prises en compte :

— Une liste de lieux géométriques pour chaque ville appartenant au département du Doubs, obtenue auprès du SDIS 25 ;
— Une liste de villes regroupées en 17 CA pour le département du Doubs ;
— Une liste des interventions de 2006 à 2018 du SDIS 25. Elle a été organisée en un ensemble de données, où chaque ligne, représentant une journée, comprend le nombre d'interventions sur une zone géographique donnée.

De la première source, on a extrait les polygones qui décrivent chaque ville. Ensuite, ils ont été regroupés par CA en tenant compte de la deuxième source. La troisième source comporte 5 versions : les données réelles et les données anonymes (4 au total avec des f variant en $[0{,}2; 0{,}4; 0{,}6; 0{,}8]$, qui garantissent ϵ_∞ en $[4{,}394; 2{,}773; 1{,}695; 0{,}811]$). Pour les deux types d'ensembles de données, on a ajouté des informations temporelles telles que l'année, le mois, le jour, le jour de la semaine, le jour de l'année, des valeurs (1 pour "Oui", 0 pour "Non") pour indiquer les années bissextiles, le premier ou le dernier jour du mois, et le premier ou le dernier jour de l'année comme attributs.

Dans la plupart des cas, les données anonymisées décrivent un nombre d'interventions plus élevé que le nombre réel. Afin de préserver l'intégrité des données, un filtre est appliqué à chaque ensemble anonymisé. Par exemple, un ensemble de données anonymisées spécifique est pris ; pour chaque ville qu'il contient, un ratio est obtenu. Ce ratio r est le résultat de la division de la moyenne du nombre d'incidents survenus l'année précédente (2017) par le jeu de données réel et le jeu de données anonymes, selon la ville. Ainsi, le nouveau nombre d'interventions anonymisées dans chaque point de données d'une ville est le résultat de la division du nombre d'interventions anonymisées par leur ratio calculé respectif.

Les données sont considérées comme séquentielles dans chaque ensemble de données. La cible est un vecteur, où chaque position et chaque valeur représentent respectivement la CA et le nombre de ses interventions, pour l'heure suivante $(t + 1)$ d'un échantillon actuel (t). Un échantillon actuel est composé du nombre actuel d'interventions dans chaque CA et des variables temporelles à ce moment. Comme la base de données fournie par le SDIS25 contient des informations sur les interventions suivies de 2006 à 2018, les modèles sont formés en utilisant les années 2006-2017 et testés en 2018.

6.2 Modélisation

Afin de faire une prévision multiple du nombre d'interventions par CA, une régression multicible est utilisée pour résoudre cette tâche. Ainsi, le "MultiOutputRegressor" de la bibliothèque scikit-learn [12] est appliqué. À cet égard, un régresseur par cible (CA) est ajusté en utilisant le régresseur XGBoost avec le paramètre $objective = `count : poisson'$ et le reste par défaut.

Six modèles ont été construits. Deux modèles formés avec les données réelles : un modèle de base qui décrit le nombre moyen d'interventions par CA pour chaque jour de la semaine ; et un second modèle construit avec XGBoost qui

prédit le nombre d'interventions par CA pour une journée entière. En outre, quatre modèles ont été construits avec des données anonymes en tenant compte de différents niveaux de garanties de confidentialité en utilisant également XGBoost.

L'hypothèse retenue est la suivante : les pompiers communiquent les données anonymes et les informations sur le ratio r de l'année précédente à des tiers afin de construire des modèles appropriés. L'efficacité des modèles est ensuite évaluée en comparant les résultats avec les données réelles de 2018, non apprises.

6.3 Résultats

Les modèles sont évalués à l'aide des mesures de l'erreur quadratique moyenne (EQM) et de l'erreur absolue moyenne (EAM). De plus, comme il s'agit d'un scénario à sorties multiples, les scores de chaque cible sont moyennés avec une moyenne uniformément pondérée sur les sorties.

Le tableau 2 présente les résultats des mesures pour une prédiction de base, pour les modèles formés avec les données originales et pour les modèles formés avec des données anonymes. Pour les ensembles de données anonymisées, les résultats sont présentés dans les deux cas où le ratio r est utilisé pour normaliser le nombre d'interventions par CA en fonction de l'année 2017 et dans celui où il ne l'est pas. Dans la figure 4, les meilleurs résultats de prédiction sont illustrés pour chaque CA en comparant le nombre initial d'interventions avec les modèles formés avec les données brutes et anonymes ($f = 0{,}60; \epsilon_\infty = 1{,}69$) pour une seule journée.

Modèle	Ratio normalisé		Ratio non normalisé	
	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 2 – Erreurs de prédiction du nombre d'interventions journalières par CA en 2018 en utilisant des données originales et anonymes normalisées et non normalisées.

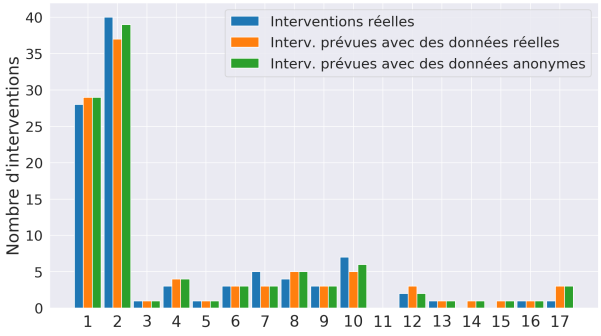

FIGURE 4 – Comparaison entre les nombres réels et prévus d'interventions par CA pour une seule journée.

6.4 Discussion

Étant donné une mise en œuvre d'un mécanisme de confidentialité différentielle locale pour la collecte de données sur la localisation des interventions, cette recherche a mis en œuvre un algorithme d'apprentissage automatique pour prédire le nombre d'interventions par CA. En comparaison avec la littérature, ce travail introduit un modèle de prévision d'interventions par période pour plusieurs CA plutôt que pour l'ensemble du département, ce qui est une tâche plus difficile. De plus, il est remarquable de constater l'amélioration du score avec les modèles formés pour une tâche aussi complexe plutôt que de développer un simple modèle de prévision comme la moyenne supposée dans cet article.

Comme on peut le constater dans le tableau 2 et dans le figure 4, les modèles formés avec des données anonymisées et normalisées peuvent également garantir une bonne utilité des données à des fins de prédiction. Il convient de noter l'utilisation d'un ratio r pour normaliser le nombre d'interventions par CA et par jour, où dans ce cas la performance de prédiction n'a pas trop diminué par rapport au modèle formé avec les données brutes. En revanche, pour les ensembles de données non normalisés, les résultats diminuent très rapidement à mesure que la garantie de confidentialité est appliquée, et après $f = 0{,}4$, les mesures de le EAM et de le EQM sont toutes deux pires que le modèle de base (la moyenne).

Les nombres en gras dans le tableau 2 représentent le meilleur résultat pour l'ensemble de données anonymisées maximisant le compromis vie privée/utilité garantissant un $\epsilon_\infty = 1{,}69$. De meilleurs résultats ont été obtenus (par exemple avec $f = 0{,}05$ et $f = 0{,}15$), cependant, les deux ensembles ont un niveau de garantie de confidentialité inférieur : $\epsilon_\infty = 7{,}33$ et $\epsilon_\infty = 5{,}02$ respectivement.

Par conséquent, dans la figure 4, il est montré pour un jour donné du mois de mars 2018 la comparaison du nombre réel et prévu d'interventions par CA en utilisant les données brutes et anonymisées ($f = 0{,}60$,$\epsilon_\infty = 1{,}69$). Avec un tel résultat, la prédiction du nombre d'interventions par CA pour le lendemain, les pompiers peuvent se préparer efficacement pour des scénarios à court, moyen et long terme. En particulier, sachant que certaines CA sont plus susceptibles de connaître des incidents, les pompiers peuvent mieux répartir les ressources humaines et matérielles ainsi que planifier la construction de nouvelles casernes. Cette approche de prédiction peut aider l'expérience humaine à prendre des décisions réelles.

Pour l'instant, un bon modèle serait celui qui peut prévoir le nombre d'interventions quotidiennes pour chaque région avec une marge d'erreur de ± 2. À l'avenir, il serait nécessaire de développer différents modèles de prévision pour chaque CA et d'établir différentes marges d'erreur en tenant compte du nombre moyen d'interventions quotidiennes. Par exemple, les deux premiers CA de la Figure 4, qui ont normalement un nombre élevé d'interventions, pourraient maintenir ± 2. Et les autres, qui ont un faible nombre d'interventions par jour, avec ± 1.

7 Conclusion

La confidentialité différentielle locale est une approche efficace utilisée pour protéger la vie privée d'une personne lors du processus de collecte de données. Plutôt que de faire confiance à un conservateur de données pour stocker les données brutes et les rendre anonymes en réponse aux requêtes (comme l'approche générale de confidentialité différentielle), la CDL permet aux utilisateurs de rendre leurs propres données anonymes avant de les envoyer au serveur du collecteur de données.

Dans cet article, l'application d'un mécanisme de CDL pour la collecte de données à des fins de préservation de la vie privée sur les lieux d'intervention des pompiers est présentée. Comme le montrent les résultats, le mécanisme "Basic One-Time RAPPOR" peut acquérir de manière adéquate des statistiques avec un bon niveau de garanties de confidentialité.

En outre, il est possible de prévoir le nombre d'interventions par CA avec des données anonymes aussi précisément qu'avec les données brutes. Ces prédictions peuvent être utilisées pour développer des outils prédictifs. Par exemple, les prévisions à court terme permettent d'optimiser les effectifs pour la semaine à venir. A moyen terme, elles permettent de redéployer de façon saisonnière les ressources matérielles et humaines dans les casernes existantes. Enfin, à plus long terme, elles peuvent aider à choisir l'emplacement géographique des futures casernes.

Nous prévoyons ainsi d'étudier dans un futur proche des approches permettant d'augmenter le volume données de manière fictive mais réaliste. L'influence de l'étape d'agrégation sur la précision des prédictions n'est pas évaluée dans cet article, qui se concentre davantage sur la partie d'anonymisation par confidentialité différentielle locale. Cette partie sera aussi adressée en travaux futurs. Enfin, la généralisation à d'autres territoires est une perspective pratique que nous travaillerons.

Remerciements

Ce travail a été soutenu par le projet CADRAN de la Région Bourgogne Franche-Comté, par l'école doctorale EIPHI-BFC (contrat "ANR-17-EURE-0002"), par le projet Interreg RESponSE et par la brigade de pompiers SDIS25.

Références

[1] S. Cerna, C. Guyeux, H. Hwang Arcolezi, A. D. P. Lotufo, R. Couturier, and G. Royer. Long short-term memory for predicting firemen interventions. In *6th International Conference on Control, Decision and Information Technologies (CoDIT 2019)*, Paris, France, apr 2019.

[2] Jean-François Couchot, Christophe Guyeux, and Guillaume Royer. Anonymously forecasting the number and nature of firefighting operations. In *Proceedings of the 23rd International Database Applications & Engineering Symposium on - IDEAS19*. ACM Press, 2019.

[3] T. T. Dang, Y. Cheng, J. Mann, K. Hawick, and Q. Li. Fire risk prediction using multi-source data : A case study in humberside area. In *2019 25th International Conference on Automation and Computing (ICAC)*, pages 1–6, Sep. 2019.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38, 1977.

[5] Cynthia Dwork. Differential privacy : A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[6] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. *J. Priv. Confidentiality*, 7(3) :17–51, 2016.

[7] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4) :211–407, 2014.

[8] 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.

[9] Christophe Guyeux, Jean-Marc Nicod, Christophe Varnier, Zeina Al Masry, Nourredine Zerhouny, Nabil Omri, and Guillaume Royer. Firemen prediction by using neural networks : A real case study. In *Advances in Intelligent Systems and Computing*, pages 541–552. Springer International Publishing, August 2019.

[10] Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C. Pierce, and Aaron Roth. Differential privacy : An economic method for choosing epsilon. In *Proceedings of the 2014 IEEE 27th Computer Security Foundations Symposium*, CSF '14, pages 398–410, Washington, DC, USA, 2014. IEEE Computer Society.

[11] Jong Wook Kim, Dae-Ho Kim, and Beakcheol Jang. Application of local differential privacy to collection of indoor positioning data. *IEEE Access*, 6 :4276–4286, 2018.

[12] 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.

[13] Latanya Sweeney. k-anonymity : A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05) :557–570, October 2002.