Forecasting the number of firefighter interventions per region with local-differential-privacy-based data

Héber H. Arcolezi, Jean-François Couchot, Selene Cerna, Christophe Guyeux, Guillaume Royer, Béchara Al Bouna, Xiaokui Xiao

Version of Record: https://www.sciencedirect.com/science/article/pii/S0167404820301619
Manuscript_24661802c33db776d96dbe5f87465677

Forecasting the Number of Firefighters Interventions per Region with Local-Differential-Privacy-Based Data

Héber H. Arcolezi[a,*], Jean-François Couchot[a], Selene Cerna[a], Christophe Guyeux[a], Guillaume Royer[b], Béchara Al Bouna[c], Xiaokui Xiao[d]

[a] *Femto-ST Institute, Univ. Bourgogne Franche-Comté, UBFC, CNRS, Belfort, France*
[b] *SDIS 25 - Service Départemental d'Incendie et de Secours du Doubs, France*
[c] *Lab., Antonine University, Hadath-Baabda, Lebanon*
[d] *School of Computing, National University of Singapore, Singapore*

Abstract

Statistical studies on the number and types of firefighter interventions by region are essential to improve service to the population. It is also a preliminary step if we want to predict these interventions in order to optimize the placement of human and material resources of fire departments, for example. However, this type of data is sensitive and must be treated with the utmost care. In order to avoid any leakage of information, one can think of anonymizing them using Differential Privacy (DP), a safe method by construction. This work focuses on predicting the number of firefighter interventions in certain localities while respecting the strong concept of DP. A local Differential Privacy approach was first used to anonymize location data. Statistical estimators were then applied to reconstruct a synthetic data set that is uncorrelated from the users. Finally, a supervised learning approach using extreme gradient boosting was used to make the predictions. Experiments have shown that the anonymization-prediction method is very accurate: the introduction of noise to sanitize the data does not affect the quality of the predictions, and the predictions faithfully reflect what

*Corresponding author
Email addresses: heber.hwang_arcolezi@univ-fcomte.fr (Héber H. Arcolezi), jean-francois.couchot@univ-fcomte.fr (Jean-François Couchot), selene_leya.cerna_nahuis@univ-fcomte.fr (Selene Cerna), christophe.guyeux@univ-fcomte.fr (Christophe Guyeux), guillaume.ROYER@sdis25.fr (Guillaume Royer), bechara.albouna@UA.EDU.LB (Béchara Al Bouna), xkxiao@nus.edu.sg (Xiaokui Xiao)

happened in reality.

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

---

1. Introduction

Emergency medical transport includes the various services useful for transporting injured people from their homes or from the place of accident to the hospital best able to take care of the patient. How this emergency medical transport is implemented depends on the country being considered, its history, and the healthy choices that have been made in the past. It usually includes the hospitals' own transport services, and often also private specialized operators (licensed private ambulance drivers). It may also include other public services, such as fire brigades. In France, for example, the latter is not only responsible for extinguishing fires, but it is also written into their status that they must take charge of part of these emergency medical transports, and this burden represents more than 80% of their activity.

This structuring has worked well in the past, nevertheless, in France as in various other countries, we have been facing a major crisis in emergency medical transport for some time now, for various reasons. The aging of the population in Western countries and the fact that older people need assistance more often leads to a higher demand for transport. The indebtedness of countries and the major economic crisis of the past decade have pushed their governments to further rationalize social spending, taking measures such as closing small centers or moving to ambulatory care (patients must be sent home as soon as possible to reduce the number of beds to be managed). However, the closure of small centers not only leads to saturation of large emergency centers but also increases the distances to be covered by health transporters. Similarly, ambulatory care increases the risk of re-hospitalization, and therefore the return journey between the hospital and the patient's home. The economic model of private ambulance drivers is only viable if the "guard" part is weak in front of the

planned medical transport (excluding emergencies). These and other elements are therefore leading to an emergency health transport crisis in various parts of the world.

One of the solutions envisaged to relieve the pressure on these transporters is to optimize the use of their resources, in order to strengthen teams during peak periods, while reducing them during off-peak ones. Such optimizations are usually implemented by asymmetric day and night staffing, and sometimes by the distinction between working days and weekend. But the crisis situation is such that it is now necessary to go much further in these optimizations, which requires a relatively clear vision of short-, medium- and longer-term needs. This prediction is possible to some extent since this emergency medical transport activity is directly related to human activity: the latter is reduced at night as people sleep, so there are fewer accidents, and the need for transport is consequently lower at night (hence the reduced shifts). However, we could go much further, considering that the activity changes according to the seasons (falling pedestrians on ice patches in winter, drowning in swimming pools in summer...), holidays, days in the week, the occurrence of planned events such as festivals or events, etc.

Not only the intervention flow can therefore probably be predicted, but also its type and location. Indeed, predicting the number of interventions per location can reduce the time required to arrive at the accident site. For example, in megacities and during heat waves, areas that are both heavily polluted (such as bus stations) and have a high density of people at risk of respiratory problems (the elderly, whose geographical distribution is known from national statistical and demographic institutes) are clearly sensitive, and pre-positioning an ambulance at these locations allows for faster action in the event of an emergency of the respiratory distress type. Reducing the time taken to arrive at the scene of the accident has material, human and economic benefits. Material, first of all, because it is possible to redeploy resources when an overload of interventions in a given area is expected (for example, due to flooding of certain rivers): with visibility, it is possible to optimize the use of current resources. Also, arriving

on the spot as quickly as possible is crucial in the case of fires, and arriving at the beginning of the fire makes it possible to limit the damage and save property and buildings. This prediction also makes it possible to optimize the use of human resources, but also to save lives in situations such as cardiac arrest and drowning, for which every second count. And these optimizations translate into economic benefits, both because of the safeguarding of property, and because premature death has a significant societal cost.

The increase in water levels following heavy rains leads to flooding events near rivers, involving personal rescue. High altitude roads have a higher risk of snow cover in winter than low altitude ones, increasing the risk of road accidents, etc. This is why some authors have recently sought to exploit artificial intelligence techniques [1, 2, 3], based on features conditioning human activity (meteorological variables, road traffic information, epidemic monitoring, etc.), in order to predict future demand in emergency medical transport. However, to be supervised, automatic learning requires the ability to put the number of interventions over the period (time, day...) for which we have these explanatory variables. In other words, it is necessary to have access to the intervention flow of the operators whose load we are trying to predict (private ambulance drivers, firefighters, etc.). The latter usually have neither the human and material resources nor the competence to deploy artificial intelligence-based solutions and are therefore obliged either to transmit this data to a trusted third party with this capacity or to release their data so that the academic world or private operators can propose ad hoc machine learning solutions.

The release of this data is therefore of undeniable interest and could help to provide solutions to the emergency health transport crisis. But this release of intervention flows is in turn problematic. First of all, it is personal data, and various legal frameworks naturally block its disclosure. Then, it is sensitive data, linked to accidents, to the rescue of people, possible deaths. As health carriers work on a just-in-time and urgent basis, human or organizational errors are always possible, which can be serious consequences, lead to lawsuits, etc. This is why these data, which were recently released in order to see predictive

tools appear, were released after anonymization. In France, for example, we recently had two publications of such flows on data.gouv.fr, a government site dedicated to such initiatives, in an open data approach. The first concerns the 2007-2017 interventions of the Service Départemental d'Incendies et de Secours de Saône-et-Loire (SDIS 71), containing the number of interventions by type and by municipality [4], while the second concerns the same types of data for SDIS 91 (Essonne department) for the period 2010-2018 [5]. In each case, anonymization was done by aggregation: monthly for the first set of data, and weekly for the second.

While the intention of these SDIS is laudable, the way in which they released these data poses two problems: the anonymization achieved is both too strong and too weak. Too strong, first of all, because carrying out one aggregation per month results in the loss of all useful information, and summarizes the interventions at a cloud of 120 points (12 per year), for which only a simple linear regression remains possible: impossible to envisage machine learning with such a data set - this is true, to a lesser extent, for data aggregated weekly. Then, too low, because this aggregation per month, or per week, was done in a blind and generalized way: if some communes have a sufficiently large number of interventions, which allows a simple temporal aggregation to achieve anonymization of the data, others conversely do not have enough. In the case of monthly aggregation, for example, there are more than 600 situations where there has been only one intervention in a commune in a given month: at this level, the simple 2-anonymity [6] is no longer satisfied, and the information leakage is obvious. Such information leaks are also numerous in the case of weekly data, and anonymization has failed for both sets of data. By analyzing this file, we learn, for example, that in the commune of Ballore (FR-71220), an intervention by the fire brigade took place in August 2014. Considering that the municipality has 86 inhabitants, it would not be very difficult to find the person who received help this month.

The objective of this article is, therefore, to show that it is possible to process such flows in a way that 1) anonymity is guaranteed, and 2) correct predictions

can be made by automatic learning on these data. This is true even if the data considered have very variable spatial densities.

Data anonymization is indeed a very active field of research, and major advances such as Differential Privacy (DP) [7] make it possible to find a fair compromise between privacy and contained information. And it is now possible to preserve both the security and utility of the released data [8]. First of all, anonymization aims to protect information about each individual, when the machine learning seeks to understand general, group trends (periodicity, seasonality, etc.): these two objectives, therefore, have, a priori, no reason to be opposed. And various similarities can be highlighted in these two approaches. For example, individuals who stand out from the crowd are obviously problematic and cannot be preserved if the objective is to produce anonymized data; these individuals also pose a problem during learning and are frequently dismissed as outliers. Similarly, learning data are usually noisy, and this noise is generally non-uniform. This asymmetry in the non-informative part of the signal makes learning more complex. Conversely, the addition of uniform noise is a classic method of anonymizing data, and this addition can, in a way, smooth out the part of the learning signal that is biased by non-uniform noise.

With these elements in mind, in this article, our objective is to apply a localized version of Differential Privacy, to transform real data so that they are both properly anonymized, and useful for automatic learning. More specifically, in this local setting (namely Local Differential Privacy - LDP), each user perturbs its data before sending it to the untrusted server. LDP has been widely applied and accepted in the process of data collection. Google applies the RAPPOR mechanism [9] to collect web browsing behavior and user's settings in Chrome. Apple [10] applies LDP to collect population statistics aiming to find commonly used emojis and new words. To other application domains, [11] applied LDP for collecting indoor position data; authors in [12] proposed a variant of LDP suitable for metric spaces (e.g., location data); and [13] proposed a protocol for finding frequent items in the set-valued LDP setting.

After receiving the noisy data from LDP, the server can compute population

statistics on the sanitized dataset. These processed data are then used for learning and prediction purposes: a multi-prediction task of the number of interventions per region, with both raw and anonymized data, is then proposed. Considered approaches encompass the use of long short-term memory for the total number of interventions [2]; a multilayer perceptron for the total number of interventions again [3]; and finally the use of XGBoost over 3h time slot, one model per two important regions, and models per motive [8].

In Figure 1, a flowchart summarizes the approach proposed and implemented in this paper. First, the algorithm takes as input the raw database (presented in Section 2) and the ϵ parameter of DP (where its theoretical background is explained in Section 3). Secondly, the LDP-based mechanism is applied to anonymize each data point (location of an intervention), which is introduced as a methodology for a privacy-preserving collection of data in Section 4. Thirdly, an intuitive Statistical-Based approach is used to estimate statistics and build a synthetic dataset (non-interactive approach in DP), which is detailed in Section 6. Both the second and third steps are based on the RAPPOR mechanism presented in [9]. Finally, using an anonymized version of the dataset (synthetic), the XGBoost technique is trained and tested for the specific task of predicting the number of interventions per region (presented in Section 7). This article ends with a conclusion section, in which the contribution is summarized and intended future work is outlined.

2. Data Presentation

The database at our disposal was provided by the fire and rescue department, SDIS 25, in the region of Doubs-France. This file has information about 382,046 interventions attended by the fire brigade from 2006 to 2018 inside their department. Each intervention is recorded in a file as a line and the main attributes of this file are shown in the Table 1 with artificial information and described as follows:

- *ID* is the intervention identifier, which is used in supplementary files;

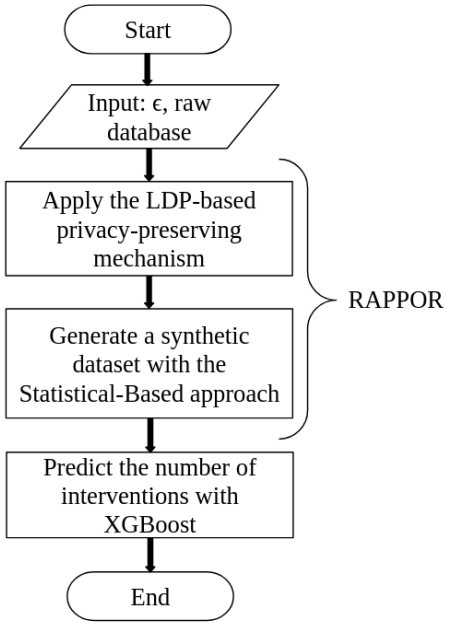

Figure 1: Flowchart of the proposed and implemented approach.

ID	SDate	Station	Town	Location
8	2008/08/08 08:08	Besançon East	Besançon	(47.2380, 6.0243)

Table 1: Main attributes of fire brigade operations data

- *SDate* is the starting date of the intervention;
- *Station* is the fire station name who attended the intervention;
- *Town* is the name of the municipality where the operation took place;
- *Location* gives the precise location (latitude, longitude) of the intervention.

Moreover, Table 2 presents data analysis of the interventions grouped *by day* in each year. The metrics are the total number of interventions (Total Interv.), the average (Average), the standard deviation (Std. Dev.), the maximum and the minimum number of interventions (Max. and Min. Interv.). As one can see in Table 2 there is a high increment in the number of interventions over

the years. That is, in 10 years the number of interventions duplicated from 17,333 in 2006 to 34,436 in 2016 and continued increasing up to 40,510 in 2018. This increment represents more work for the next years, where a better optimization of resources must be considered to continue improving response times to incidents and to better attend the population.

Year	Total Interv.	Average	Std. Dev.	Max. Interv.	Min. Interv.
2006	17,333	47	20	131	17
2007	19,277	53	16	116	23
2008	18,021	49	14	117	26
2009	28,669	79	38	257	22
2010	29,604	81	26	328	42
2011	33,645	92	39	403	48
2012	29,079	79	16	143	52
2013	29,760	82	14	145	47
2014	30,641	84	14	164	54
2015	33,518	92	17	154	57
2016	34,436	94	28	556	60
2017	37,553	102	16	165	61
2018	40,510	111	21	265	73

Table 2: Data analysis of the interventions during 2006-2018.

3. Theoretical Background on (Local) Differential Privacy

Let $\mathcal{A}$ be an algorithm used to publish aggregate information of a private database. Differential privacy (DP)[14] is as a constraint (property) on $\mathcal{A}$ which limits the disclosure of private information of records whose information is in the database. Roughly, $\mathcal{A}$ is differentially private if an observer seeing its output cannot tell if a particular individual's information was used in the computation.

Let ϵ be a positive real number which intuitively corresponds to the leakage

level. The higher the value of this variable, the more important is the information leakage. Let $\mathrm{im}\,(\mathcal{A})$ denotes the image of $\mathcal{A}$, *i.e.*, the set of all possible outcomes by $\mathcal{A}$. The algorithm $\mathcal{A}$ is said to provide ϵ-differential privacy if, for all datasets D_1 and D_2 that differ on the data of one person, and for all subsets R of $\mathrm{im}\,(\mathcal{A})$, we have

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

Intuitively, given $\Pr[\mathcal{A}(D_2) \in R]$ (the probability that a D_2 data set can be anonymized into an element of R) and given ϵ the amount of leakage. This equation gives an upper bound of the probability that a data set D_1 can be anonymized into an element of R, which is thus an information leakage.

Differential privacy allows composability (of independent mechanisms who are $\epsilon_1, \ldots, \epsilon_n$ DP...), robustness to post-processing ($F(\mathcal{A})$ is ϵ_n DP for any function F).

However, this approach requires the whole dataset to be complete, stored in a safe manner and further anonymized. Anonymisation is not done before. It is the objective of the local differential privacy introduced in [15]. In this approach, data are sanitized by the user in a probabilistic manner before sending them to the collector. A simple example is to ask a person to answer the question "Do you live in Belfort?", according to the following procedure:

Throw a coin.

- If tail, then throw the coin again (ignoring the outcome), and answer the question honestly.
- If head, then throw the coin again and answer "Yes" if head, "No" if tail.

This basic stochastic method is summarized in Figure 2. Let t_y be the proportion of truth "Yes" answers and c_y be the proportion of observed "Yes" answers. The following equation gives an estimated relation between these two variables

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

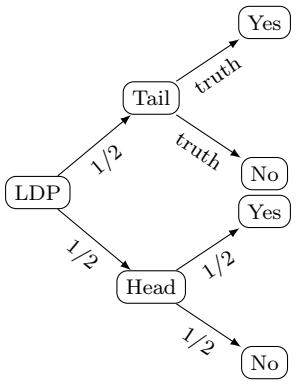

Figure 2: Summary of what is sent to the collector by basic LDP

The higher the number of experiments, the closer the proportion of random "Yes" responses will be to 1/4 and the closer the number of times the truth is told, the more accurate the estimate will be. In this case, t_y can be estimated by

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

The algorithm $\mathcal{A}$ is said to provide ϵ-local differential privacy if, for all pairs of user's possible private data v_1 and v_2 and all subsets R of im$\mathcal{A}$:

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

4. Privacy-Preserving Firemen Interventions' Location Data Collection (User side)

The first question one can ask is if an intervention is a sensitive attribute. The answer is certainly yes because the fire brigade would not have been called if the situation had not been severe enough. For example, consider the scenario where a person who habits in a small town has acquired a very particular disease. If it is known that for this period one intervention happened in this town where normally it seldom does, there is a high probability that the fire brigade intervened for this person.

Therefore, the purpose of this task is to implement a privacy-preserving mechanism to firefighters' intervention's location using the concept of local differential privacy previously described. Next, given a specific period, the challenge is to estimate the number of firemen interventions within the considered locations using the anonymized data to build a synthetic dataset. To summarize, rather than precisely determining each intervention's coordinates, the objective of this paper is hiding the information of intervention's location such that statistics on the number of interventions per location can be acquired with acceptable utility. Figure 3 illustrates an outline of the approach and it is summarized in the following.

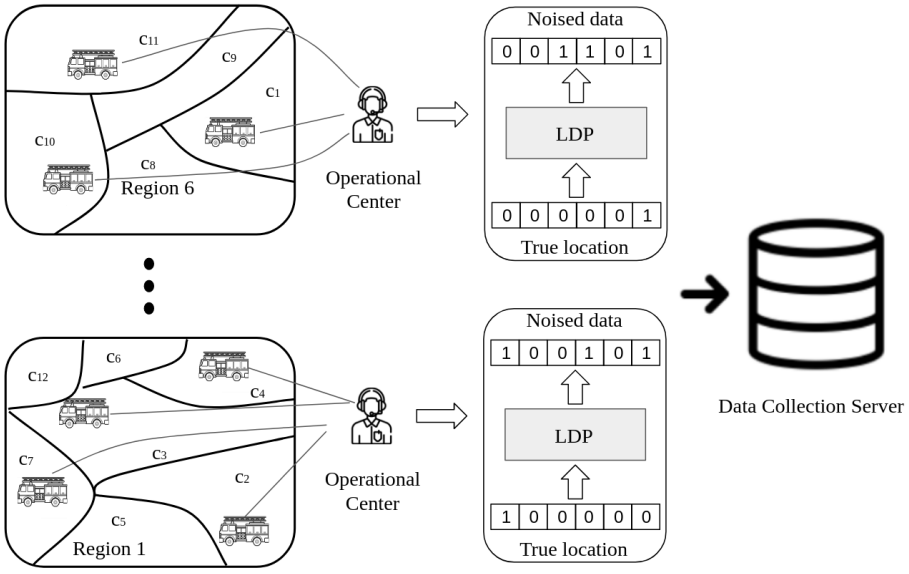

Figure 3: An outline of the approach applied to collect firemen intervention's location data preserving privacy.

In the proposed approach, the first step to guarantee the privacy of each interventions' location is grouping the towns which happened each intervention at the level of a bigger city (region) to obtain events that are sufficiently representative in number. For example, one can notice in Figure 3 that a set of $C = \{c_1, c_2, ..., c_{12}, ..., c_m\}$ small towns are grouped to $n = 6$ regions.

In this context, using the data at our disposal, 608 towns where interventions happened in the Doubs department were generalized to $n = 17$ regions using the public dataset available in [16]. The 17 regions are: (1) CA du Grand Besançon, (2) CA Pays de Montbéliard Agglomeration, (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. Figure 4 illustrates the department of Doubs with the respective towns and its agglomeration to regions.

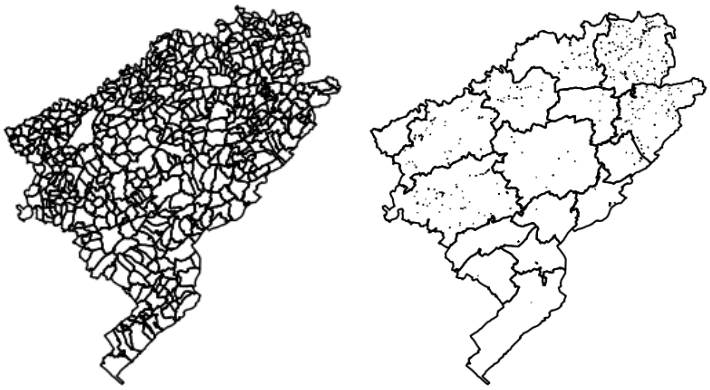

Figure 4: Towns in the department of Doubs agglomerated by regions.

Secondly, to improve the level of privacy for each intervention, the LDP-based mechanism "Basic One-time RAPPOR" introduced by [9] is applied. This algorithm is a simplification of the RAPPOR mechanism, which uses bloom filters and hash functions to map reports sent by users and it has two levels of randomized responses namely permanent and instantaneous ones.

However, in the "Basic One-time RAPPOR", it is applied just one step of randomized response using a deterministic mapping of the $n = 17$ regions into

one-hot-encoded vectors. The motivation to use this straightforward algorithm is based on two assumptions:

- The agglomerations of towns (regions) are known a priori allowing the deterministic mapping rather than using hash functions and bloom filters;
- The exact location of each intervention has unique (x, y) coordinates, which allows sending one unique report per intervention based on its big agglomeration. Notice in Figure 3 that all interventions which happened in the area of "Region 1" will report a noised location based on the same true value.

A technical application of this algorithm in our case study is described below:

1. **True location signal.** Let $R = \{r_1, r_2, ..., r_n\}$ be a set of n regions in consideration, where each subscripts represent a unique region ID. Hence, a n-bit array, B (which denotes the current intervention location) is defined as

$$B_k = \begin{cases} 1, & \text{if } k = i \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where in this case, B_k represents the value of the k-th bit in B with $k \in [1, n]$. That is, the bit corresponding to the regions ID is set to one, while the others are set to zero (as the true location in Figure 3).

2. **Permanent randomized response.** Next, each bit in B (from the previous step) is perturbed by applying the concept of randomized response as follows:

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

where f is a probability value between 0 and 1, which controls the level of ϵ-differential privacy guarantee (see [9] for mathematical proofs). One

can notice the direct relationship between privacy and utility by varying f where increasing it guarantees more privacy with the cost of adding more noise from B to U.

3. **Final report.** The permanent randomized response B is transmitted to the data collector server.

The ϵ-local differential privacy level has been shown in [9] to be at worst ϵ_∞ defined as

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

The original proof contains steps that are not easy to follow. The Appendix A presents another proof for this value.

4.1. Example

Assume that an intervention occurred in the area of r_3, which represents the 3-rd region from $n = 8$ ones. Therefore, its true location signal B is described as follows:

$$B = [0, 0, 1, 0, 0, 0, 0, 0] \tag{6}$$

whereas one can see that the 3-rd bit of B is set to one. At this step, the privacy guarantee of the intervention location is ensured by the agglomeration area, however, in several scenarios, an attacker could make use of background knowledge and external sources to infer the exact location (in this case, the town). Hence, applying the Equation (4) with, e.g., $f = 0.3$, one possible permanent response U is as follows:

$$U = [1, 0, 1, 0, 0, 1, 0, 0] \tag{7}$$

where given random properties depending on f, both 1-st bit and 6-th one are also set to one.

Therefore, as one can see, the location information is no longer easy to be discovered, as the concept of LDP ensures that any true region (input) could have generated the noised output U with a bounded probability ratio of e^{ϵ_∞}.

5. Generating a synthetic dataset (Server side)

Considering a specific study period, the objective is to estimate the number of interventions per location associated with the i-th region, r_i. In this context, a synthetic dataset can be built with this estimation, which is considered as a non-interactive case of DP. More specifically, this dataset is generated by statistics using only anonymous location data and it is released just one time for all other intended tasks.

Hence, within a specific time, let $set(U)$ be a set of permanent randomized responses and $set(B)$ be the corresponding set of original location bit arrays. Further, assume that $|set(U)|$ and $|set(B)|$ denote the number of elements in each respective set. Naturally, $|set(U)| = |set(B)|$.

Therefore, the estimated number of interventions $NBint_{est}$ per region location r_i for $i \in [1, n]$ is acquired by a Statistical-Based (SB) approach as follows [9]:

$$NBint_{est}(r_i) = \frac{1}{1-f} \cdot \left(N_i - \frac{f \cdot N_{total}}{2} \right) \tag{8}$$

where N_{total} is the number of permanent randomized responses $|set(U)|$ and N_i is the total number of permanent randomized responses whose i-th bit is set to 1. It is noteworthy that Equation (8) can estimate negative numbers, hence, the $max(0, NBint_{est})$ function is used.

To evaluate the SB result, the density estimation of an i-th region location associated with r_i is calculated as follows [11]:

$$Density_{est}(r_i) = \frac{NBint_{est}(r_i)}{\sum_{y=1}^{n} NBint_{est}(r_y)} \tag{9}$$

where n is the number of region, and, hence, the error rate (ER) metric is defined as:

$$ER = \frac{1}{n} \sum_{i=1}^{n} |Density_{actual}(r_i) - Density_{est}(r_i)| \tag{10}$$

where $Density_{actual}(r_i)$ and $Density_{est}(r_i)$ correspond to the actual and estimated density, respectively, of the region associated with the i-th location. Rather than calculating the root mean squared error over the estimated and actual number of interventions, the error rate is calculated over the density value motivated to normalized values between 0 and 1.

6. Anonymization Experiments

To evaluate the proposed approach of anonymizing firemen intervention's location, several simulations are performed with different values of f, which determines the level of ϵ_∞-differential privacy. In the experiments, f will vary in $[0.1, 0.2, ..., 0.8, 0.9]$, which guarantees ϵ_∞-differential privacy between $[5.89, 4.39, ..., 0.81, 0.4]$.

Hence, using the statistical-based approach (Equation (8)), the objective is estimating the number of interventions per region considering different scenarios of time. The scenarios of time are described in the following. The first one to analyze is with one-year data (13 data points), which allows at the beginning of a year the fire brigade to better distribute their budget around its centers according to the number of interventions per region. Next, a one-month scenario (156 data points) is considered. And, similar to before, the fire brigade can have high-utility statistics from a third-party company to reorganize budgets and personnel each month. Lastly, a one-day scenario (4748 data points) is taken into consideration such that machine learning tasks could be applied in this amount of data.

These experiments will allow evaluating the relationship between ER versus data size (period of analysis) according to ϵ_∞ in order to find the best privacy-utility trade-off for different applications. Each scenario allows the fire brigade to have an anonymous database of intervention's locations where third party companies or the human resources department itself could acquire high-utility statistics. More specifically, synthetic datasets will be build based on the SB approach, which will contain the number of interventions per region for each

scenario of time.

6.1. Results

For the sake of brevity, considering only three values for $\epsilon_\infty = [5.89, 2.19, 0.40]$ (resp. $f = [0.1, 0.5, 0.9]$), Table 3 presents the following metrics: the Average ER (ER Av.), the ER Standard Deviation (ER. Std.), the minimum (Min. ER) and maximum (Max. ER) errors, for each scenario of time. That is, as statistics are acquired, for example, for each year, the error will be summarized at once (considering all years) in Table 3.

To better illustrate the results from Table 3, Figure 5 shows the relationship of ER and ϵ_∞ for: each year (2006-2018), with zoom for the last 8 months of 2018, and with zoom for the last 8 days of December 2018, respectively. Moreover, Figure 6 illustrates the statistics acquired on the number of interventions for the year 2013, the first month of 2017, and a precise day in January 2016, with the three values for $\epsilon_\infty = [5.89, 2.19, 0.40]$ ($f = [0.1, 0.5, 0.9]$ a low, a medium, and a high privacy guarantee). All three specific dates were chosen at random for illustration purposes.

ϵ_∞	**Scenario**	**ER Av.**	**ER Std.**	**Min. ER**	**Max. ER**
5.89	One-year	0.001209	0.000271	0.000894	0.001750
	One-month	0.004045	0.001081	0.002054	0.007426
	One-day	0.017675	0.005265	0.005115	0.048845
2.20	One-year	0.003992	0.000922	0.002116	0.005644
	One-month	0.012813	0.002920	0.006475	0.021311
	One-day	0.042584	0.010536	0.014006	0.092509
0.40	One-year	0.018785	0.003726	0.012430	0.024008
	One-month	0.043107	0.010174	0.022024	0.070537
	One-day	0.077103	0.015029	0.029918	0.117647

Table 3: Metrics results for comparing the ER in different scenarios of time and ϵ_∞-differential privacy.

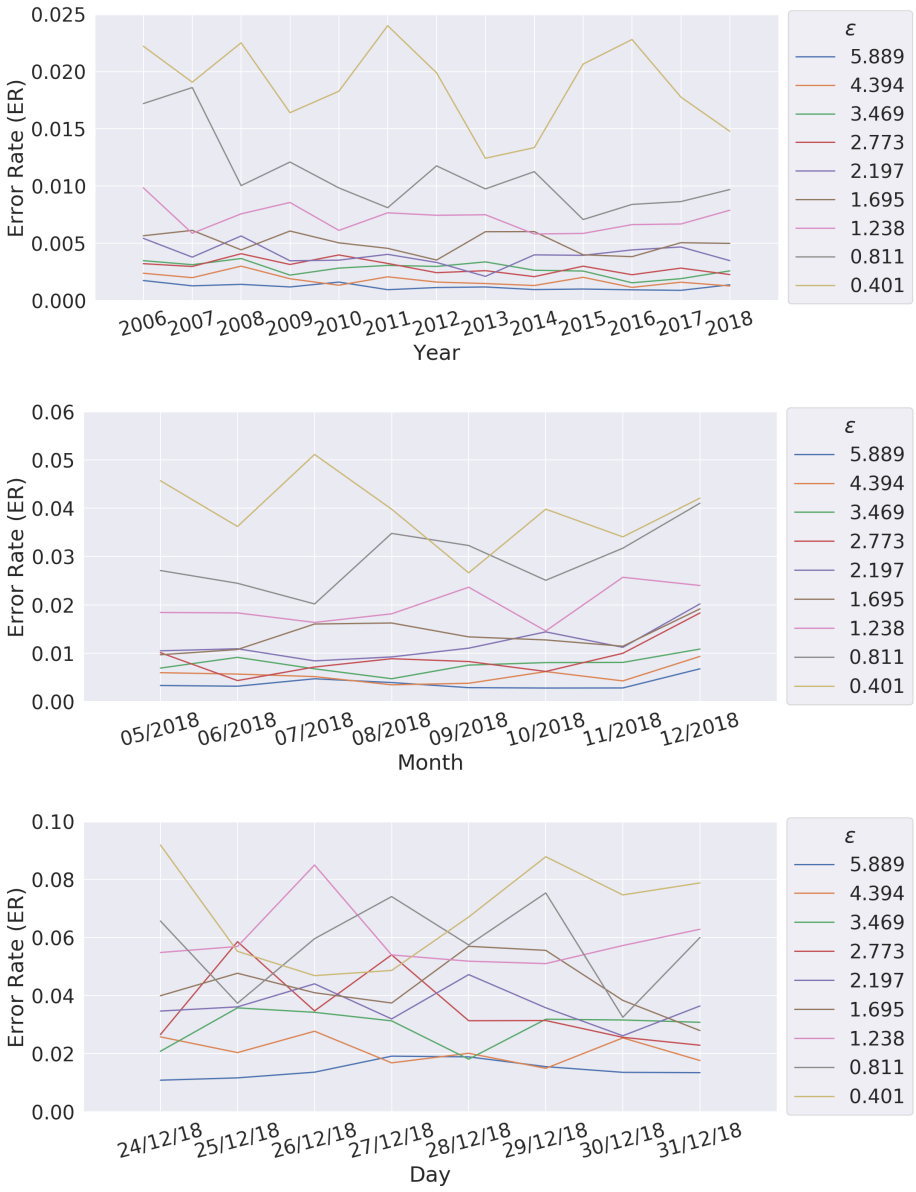

Figure 5: Comparison between error rate and period of analysis (data size) varying ϵ_∞.

6.2. Discussions

As one can notice in Table 3 and Figures 5 and 6, the LDP-based mechanism can be well applied to the collection of firemen interventions' location for the

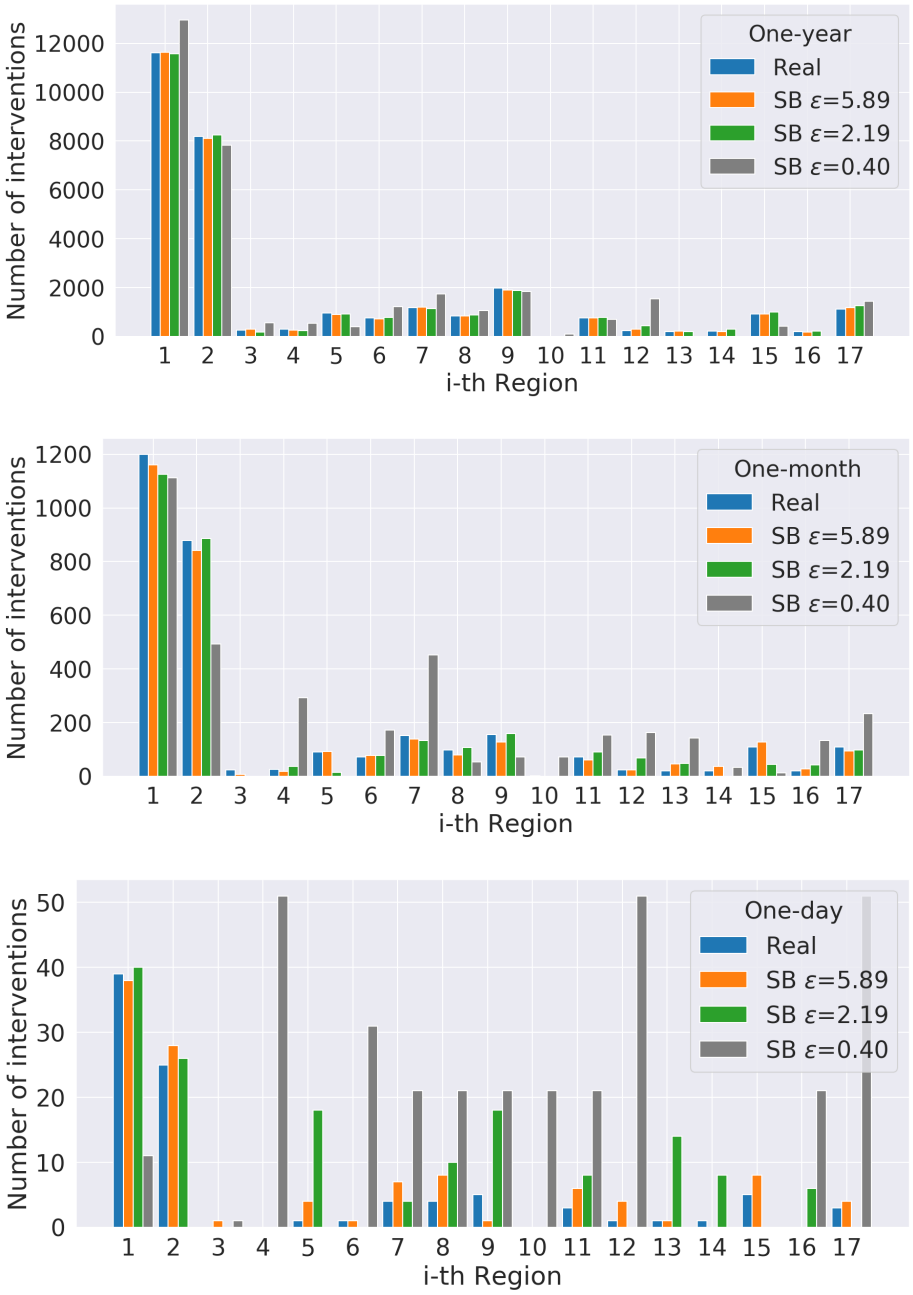

Figure 6: Analysis between the real and estimated number of interventions per region.

purpose of inferring the number of interventions per region. As LDP ensures the privacy of individuals by perturbing the data before sending it to the data collector, in this case, the fireman responsible to report interventions will apply the perturbation to the location of interventions before sending it to the data server (as Figure 3 illustrates).

It is noteworthy that the ER decreases as the data size increases. This is due to the LDP setting, which requires a big amount of data to guarantee a good balance of noise. For example, for a one-year analysis, the number of interventions is at least 17,333 in 2006, while the average per day is just 47 for the same year. For this reason, the utility of the data decreases for small-time scenarios as both one-month and one-day cases presented in this paper. Hence, one has to balance the application of the anonymized data. For instance, if one intends to acquire statistics per year, results are very accurate with good privacy guarantees. However, if one intends to apply machine learning tasks to this data (as presented in the next section), a one-day scenario is more appropriate but with higher error.

Moreover, the relationship between ER and privacy guarantees is natural, whereas the ϵ_∞-differential privacy guarantee is enhanced, more noise is added to the data and the utility of it decreases. However, as aforementioned, the data size influences very much at this step. One can see in Table 3 that while the 'ER. Av.' for the one-year analysis is around $1e-3$ for the first two levels of ϵ_∞-DP guarantee, it is not the case for the one-day scenario with the same metric between $1e-2$ and $4e-2$. The one-month scenario is the middle-low scenario with reasonable data points but not sufficient as the one-year case to provide good results. In this case, reasonable metrics are acquired in comparison with the one-day scheme.

Figure 5 summarizes both relationships of ER with data size as the level of ϵ_∞-differential privacy guarantee increases (smaller ϵ provides strong privacy guarantees). While the one-year scenario with the maximum level of privacy guarantees has ER around 0.02, the one-month scenario achieves this ER for both last values of privacy guarantees and the one-day scenario reaches this ER

already with the second-lowest privacy guarantee. Furthermore, in Figure 5 one can see the statistics acquired on the number of interventions for each period, where small errors are acquired for the one-year case and considerable ones for both one-month and one-day schemes.

Therefore, as also highlighted in the literature, the choice of ϵ depends on several factors (data size, the application domain) and one has to appropriately balance it considering the privacy of users and utility of data. In our case, as 608 towns were generalized to $n = 17$ regions, privacy could be slightly decreased to acquire good utility for generating statistics (e.g., with $\epsilon_\infty = 5.89$ as presented in Figure 6). In the literature, common values to ϵ are within the range $0.01 - 10$ [17]. In the original paper of RAPPOR [9], authors experimented $(f, q, p) = (0, 0.75, 0.5)$ to non-longitudinal data (sent just once), which guarantees $\epsilon_1 = 1.09$; $(f, q, p) = (0.75, 0.75, 0.5)$, which guarantees $\epsilon_\infty = 2.05$ and $\epsilon_1 = 0.53$ for collecting Google Chrome homepages (with approximately 14 million reports); and $(f, q, p) = (0.5, 0.75, 0.5)$, which guarantees $\epsilon_\infty = 4.39$ and $\epsilon_1 = 1.07$. In [11], authors used $(f, q, p) = (0.2, 0.75, 0.25)$, which provides $\epsilon_\infty = 4.39$ and $\epsilon_1 = 1.69$ for collecting indoor positions using real data.

7. Forecasting Firemen Interventions per Region

The purpose of this task is to implement a state-of-the-art machine learning algorithm, namely the extreme gradient boosting (XGBoost), for forecasting the number of interventions per day of the $n = 17$ regions in Doubs-France. As the main objective, anonymized files will be used to construct models in the interest of evaluating the utility of the data with different levels of ϵ_∞-differential privacy in comparison with the original one.

7.1. Data preparation

Three initial sources were considered:

- A list of geometric locations with map projection epsg:2154 for each town belonging to the Doubs department, obtained from the SDIS 25.

- A list of towns grouped in 17 regions for the Doubs department. The file was extracted from the public dataset available in [16].

- A list of interventions from 2006 to 2018, shared by the SDIS 25. It was organized in a dataset, where each row, representing one day, comprises the number of interventions per region. As shown in the previous section, statistics on the number of interventions per day can be acquired with an acceptable margin of error, which has sufficient data points (4748).

From the first source, it was extracted the polygons that describe each town. Then, they were grouped by region considering the second source. Thus, it is obtained a final list with the new polygons for each region as illustrated in Figure 4.

The third source has 10 versions: the real data and the 9 other anonymized ones from following LDP as described in Section 6 (where $f \in [0.1, 0.2, ..., 0.8, 0.9]$, *i.e.* which guarantees ϵ_∞-differential privacy between $[5.89, 4.39, ..., 0.81, 0.4]$). For both types of datasets, it was added temporal information such as year, month, day, weekday, year day, values (1 for 'yes', 0 for 'no') to indicate leap years, first or last day of the month, and first or last day of the year as attributes.

Due to the $max(0, NBint_{est})$ function, in most cases, the anonymized data describe a higher number of interventions than the real one. In order to keep the data integrity, a filter is applied to each anonymized set. As an instance, a specific anonymized dataset is taken; for each town contained in it, a ratio is obtained. The ratio is the result of dividing the means of the number of incidents happened in the previous year (2017) from the real dataset and the anonymized one, according to the town. Thus, the new number of anonymized interventions in each data point of a town is the result of dividing again the number of anonymized interventions by their respective calculated ratio.

The data are considered sequential in each dataset. The target is a vector, where each position and value represent the region and the number of its interventions respectively, for the next hour $(t+1)$ of a present sample (t). A present

sample is composed of the present number of interventions in each region and the temporal variables at that time. As the database provided by SDIS25 has information about interventions attended from 2006 to 2018, models are trained using the years 2006-2017 and tested in 2018.

7.2. Modeling

In order to make a multi-forecast of the number of interventions per region, the multi-target regression is used to solve this task. Hence, the "MultiOutputRegressor" from the scikit-learn library [18] is applied. In this regard, one regressor per target (region) is fitted using the XGBoost regressor with the parameter $objective = `count : poisson'$ and the rest as default.

Six models were built. Two models trained with the real data: one as a baseline that describes the average number of interventions in each day of the week per region; and a second one built with XGBoost that predicts the number of interventions per region for an entire day. Besides, four models were built with anonymized data considering different levels of privacy guarantees using XGBoost too.

The assumption made here is: the firemen brigade releases the anonymized data and the ratio information ("filter") from the last year to third party companies and academic institutions to build appropriated models for the real system. Hence, to evaluate the effectiveness of the models, they are all tested using the real data for 2018.

7.3. Results

The models are evaluated with the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) metrics. Moreover, as it is a multi-output scenario, the scores for each target are averaged with a uniformly weighted mean over outputs ('uniform_average') [18].

For the sake of brevity, considering only four values for $\epsilon_\infty = [4.39, 2.77, 1.69, 0.81]$ (resp. $f = [0.2, 0.4, 0.6, 0.8]$), Table 4 presents metrics results for a Baseline prediction, for models trained with the original data and

for models trained with anonymized data. For anonymized datasets, results are presented for both cases where the 'ratio' is used for normalizing the number of interventions per region according to the year 2017 or not.

Model	Normalized ratio		Non-normalized ratio	
	MAE	**RMSE**	**MAE**	**RMSE**
Baseline (mean)	-	-	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 4: Metric results for forecasting the number of interventions per region each day of 2018 using both normalized and non-normalized original and anonymized data.

Additionally, Figures 7 and 8 better illustrate the results from Table 4 in respect to the RMSE and MAE metrics with the f parameter varying from $f = 0.1$ to $f = 0.9$. In Figure 9, the best prediction results are illustrated for each region comparing the original number of interventions with models trained with the raw and anonymized data ($f = 0.60; \epsilon_\infty = 1.69$) for one single day of March 2018.

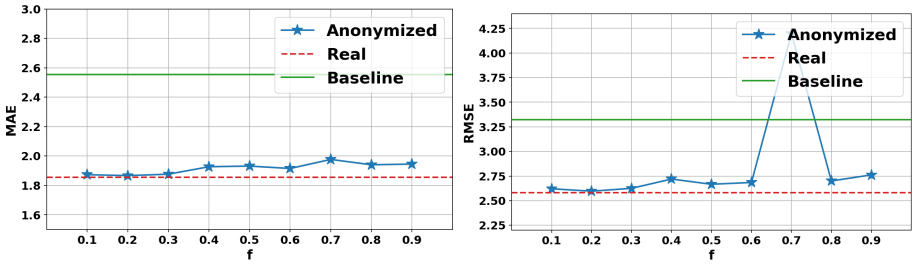

Figure 7: MAE and RMSE metrics for the normalized prediction models.

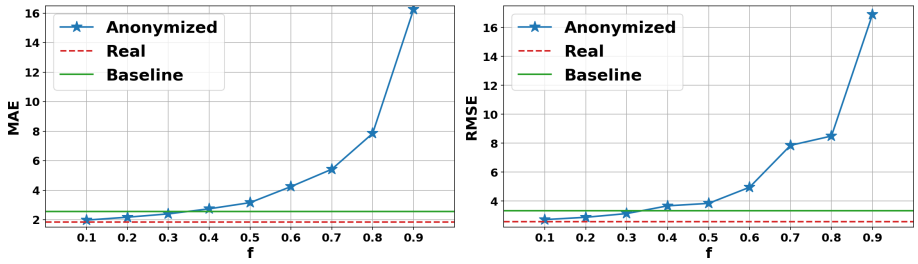

Figure 8: MAE and RMSE metrics for the non-normalized prediction models.

7.4. Discussion

Aiming to evaluate the privacy-utility trade-off given the implementation of a local differential privacy mechanism for collecting data of interventions' location, this research implements a machine learning algorithm to predict the number of interventions per region. In comparison with the literature, this work introduces a forecasting model for several regions rather than only the total number of interventions per period, which is a more difficult task. Moreover, it is remarkable the improvement of the score with the trained models for such complex task rather than developing a simple prediction model as the baseline (mean) assumed in this paper.

As one can notice in Table 4 and in Figures 7 and 8, models trained with anonymized and normalized data can also guarantee a good utility of the data for prediction purposes. It is noteworthy the use of a 'filter' to normalize the number of interventions per region and day, where in this case the prediction performance did not decrease too much in comparison with the model trained with the raw data. In contrast, for non-normalized datasets, the results decrease very fast as the privacy guarantee is enforced, and after $f = 0.4$ both MAE and RMSE metrics are worse than the baseline (mean) model.

The numbers in bold in Table 4 represent the metrics results using the anonymized dataset who has the best privacy-utility trade-off. Even though better results were found with $f = [0.1, 0.2, 0.3], \epsilon_\infty = [5.89, 4.39, 3.46]$ (as one can see in Figure 7), their privacy guarantees are too low considering a real-

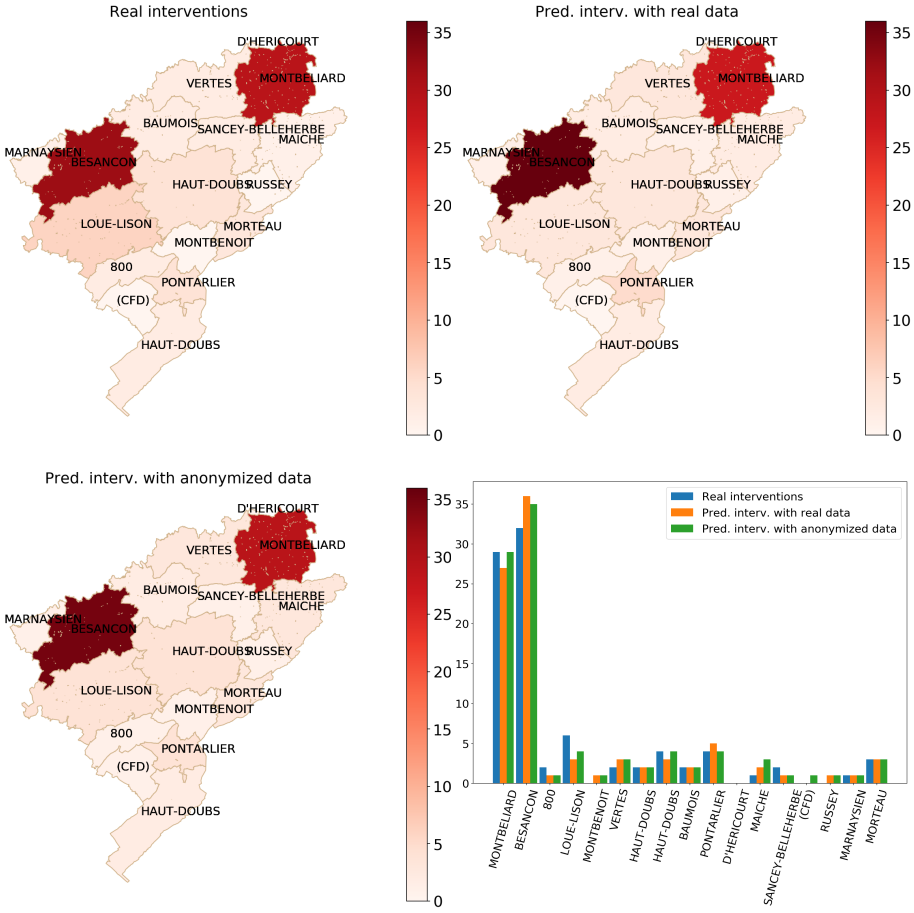

Figure 9: Comparison of the real and predicted number of interventions per region for a single day.

world application. Moreover, in our analysis, even better results were found with $f = 0.05$ and $f = 0.15$; however, both has even lower privacy guarantees with $\epsilon_\infty = 7.33$ and $\epsilon_\infty = 5.02$ respectively (as higher ϵ_∞ represents more leakage of information in DP theory).

Hence, in Figure 9, it is shown for a given day of March 2018 the comparison of the real and predicted number of interventions per region using the raw data and best-version of anonymized data ($f = 0.60, \epsilon_\infty = 1.69$). With such a result, the prediction of the number of interventions per region for the next day, the

fire brigade can efficiently prepare themselves for short-, middle-, and long-term scenarios. In particular, knowing that certain regions are more prospect to happen incidents, the fire brigade can better allocate the human and machinery resources as well as planning the construction of new barracks.

8. Conclusion

Local differential privacy is a state-of-the-art approach used to protect an individual's privacy in the process of data collection. Rather than trusting in a data curator to have the raw data and anonymize it to output queries (as the general Differential Privacy approach), LDP allows users to anonymize its own data before sending it to the data collector server.

In this paper, the application of an LDP mechanism for privacy-preserving collecting data purposes of firemen interventions' locations is introduced. As shown in the results of Section 6, the 'Basic One-Time RAPPOR' mechanism can adequately acquire statistics with a good level of privacy guarantees. In this case, an attacker cannot distinguish between values $v1$ or $v2$ (named B as the real locations of interventions), because both have approximately the same probability to generate the noised output (U).

Moreover, as shown in Section 7, it is possible to forecast the future number of interventions per region with anonymized data as well as with the raw data. More specifically, the work in this article shows that data flows such as emergency health transport, which is sensitive at the outset but can be very useful, can be properly anonymized in order to avoid information leakage, while remaining useful for optimization purposes. They can be used to develop predictive tools, and these tools can be used for many things. Short-term predictions would make it possible to optimize shifts for the coming week, anticipate by providing emergency reinforcement during peaks, and pre-position vehicles. In the medium term, these predictions would make it possible to redeploy seasonally the material and human resources to existing barracks, as well as to assist in holiday planning, given the expected workload in the coming months. Fi-

nally, in the longer term, such predictions, made possible by such learning from anonymized data, would make it possible to anticipate the future needs (human and material) necessary to maintain a certain quality of service, while helping to choose the geographical location of future barracks.

For future work, improvements to the multi-forecast model are planned. For instance, will be added to the dataset more explanatory variables such as meteorological and traffic data, where feature selection techniques will be used to improve the performance of models. Moreover, techniques for tuning the hyperparameters of the models will be implemented.

Acknowledgment

This work was supported by the Region of Bourgogne Franche-Comté CADRAN Project, by the EIPHI-BFC Graduate School (contract "ANR-17-EURE-0002"), by the Interreg RESponSE project, and by the SDIS25 firemen brigade.

References

References

[1] T. T. Dang, Y. Cheng, J. Mann, K. Hawick, 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), 2019, pp. 1–6. doi:10.23919/IConAC.2019.8894971.

[2] S. Cerna, C. Guyeux, H. H. Arcolezi, A. D. P. Lotufo, R. Couturier, G. Royer, Long short-term memory for predicting firemen interventions, in: 6th International Conference on Control, Decision and Information Technologies (CoDIT 2019), Paris, France, 2019. URL: https://doi.org/10.1109/codit.2019.8820671. doi:10.1109/codit.2019.8820671.

[3] C. Guyeux, J.-M. Nicod, C. Varnier, Z. A. Masry, N. Zerhouny, N. Omri, G. Royer, Firemen prediction by using neural networks: A real case study,

in: Advances in Intelligent Systems and Computing, Springer International Publishing, 2019, pp. 541–552. URL: https://doi.org/10.1007/978-3-030-29516-5_42. doi:10.1007/978-3-030-29516-5_42.

[4] Satistiques mensuelles fournies par le service départemental d'incendies et de secours (sdis 71), https://www.data.gouv.fr/fr/datasets/interventions-des-pompiers-od71/, 2013. Accessed: 2019-12-13.

[5] Données hebdomadaires sur les interventions des sapeurs-pompiers de l'essonne, https://www.data.gouv.fr/fr/datasets/interventions-des-pompiers/, 2018. Accessed: 2019-12-13.

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

[7] C. Dwork, A. Roth, et al., The algorithmic foundations of differential privacy, Foundations and Trends® in Theoretical Computer Science 9 (2014) 211–407.

[8] J.-F. Couchot, C. Guyeux, G. 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. URL: https://doi.org/10.1145/3331076.3331085. doi:10.1145/3331076.3331085.

[9] U. Erlingsson, V. Pihur, A. Korolova, Rappor: Randomized aggregatable privacy-preserving ordinal response, in: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, CCS '14, ACM, New York, NY, USA, 2014, pp. 1054–1067. URL: http://doi.acm.org/10.1145/2660267.2660348. doi:10.1145/2660267.2660348.

[10] A. Apple Differential Privacy Team, Learning with privacy at scale, 2017.

[11] J. W. Kim, D.-H. Kim, B. Jang, Application of local differential privacy to collection of indoor positioning data, IEEE Access 6 (2018) 4276–4286.

[12] M. S. Alvim, K. Chatzikokolakis, C. Palamidessi, A. Pazii, Metric-based local differential privacy for statistical applications, CoRR abs/1805.01456 (2018).

[13] T. Wang, N. Li, S. Jha, Locally differentially private frequent itemset mining, in: 2018 IEEE Symposium on Security and Privacy (SP), IEEE, 2018. URL: https://doi.org/10.1109/sp.2018.00035. doi:10.1109/sp.2018.00035.

[14] C. Dwork, F. McSherry, K. Nissim, A. D. Smith, Calibrating noise to sensitivity in private data analysis, J. Priv. Confidentiality 7 (2016) 17–51.

[15] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, A. Smith, What can we learn privately?, in: 2008 49th Annual IEEE Symposium on Foundations of Computer Science, IEEE, 2008. URL: https://doi.org/10.1109/focs.2008.27. doi:10.1109/focs.2008.27.

[16] Liste et composition 2018, https://www.collectivites-locales.gouv.fr/liste-et-composition-2018/, 2018. Accessed: 2019-12-01.

[17] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, A. Roth, Differential privacy: An economic method for choosing epsilon, in: Proceedings of the 2014 IEEE 27th Computer Security Foundations Symposium, CSF '14, IEEE Computer Society, Washington, DC, USA, 2014, pp. 398–410. URL: https://doi.org/10.1109/CSF.2014.35. doi:10.1109/CSF.2014.35.

[18] 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, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

Appendix A. Value of ϵ_∞-local differential privacy level.

Let us prove that the one-time RAPPOR algorithm $\mathcal{A}$ verifies ϵ-local differential privacy with ϵ equal to ϵ_∞ as defined in (5).

Let us thus find a bound of $\dfrac{\Pr[\mathcal{A}(v_1) \in R]}{\Pr[\mathcal{A}(v_2) \in R]}$, for all pairs of user's possible private data v_1 and v_2 and all subsets R of $\mathrm{im}(\mathcal{A})$:

$$\begin{aligned}
\frac{\Pr[\mathcal{A}(v_1) \in R]}{\Pr[\mathcal{A}(v_2) \in R]} &\leq \max_{U \in R} \frac{\Pr[\mathcal{A}(v_1) = U]}{\Pr[\mathcal{A}(v_2) = U]} \\
&= \max_{U \in R} \frac{\Pr[B^1 = U]}{\Pr[B^2 = U]} \\
&= \max_{U \in R} \frac{\prod_{k=1}^{n} P[B_k^1 = U_k]}{\prod_{k=1}^{n} P[B_k^2 = U_k]}
\end{aligned}$$

Thanks to Equation (4), it is easy to establish that $P(U_k = B_k) = 1 - \frac{f}{2}$ and that $P(U_k \neq B_k) = \frac{f}{2}$ for any k, $1 \leq k \leq n$. We thus have

$$\begin{aligned}
\frac{\prod_{k=1}^{n} P[B_k^1 = U_k]}{\prod_{k=1}^{n} P[B_k^2 = U_k]} &= \frac{\prod_{k=1}^{n} \left(1 - \frac{f}{2}\right)^{|U_k - B_k^1|} \left(\frac{f}{2}\right)^{1-|U_k - B_k^1|}}{\prod_{k=1}^{n} \left(1 - \frac{f}{2}\right)^{|U_k - B_k^2|} \left(\frac{f}{2}\right)^{1-|U_k - B_k^2|}} \\
&= \prod_{k=1}^{n} \frac{\left(\frac{2}{f} - 1\right)^{|U_k - B_k^1|}}{\left(\frac{2}{f} - 1\right)^{|U_k - B_k^2|}} \\
&= \prod_{k=1}^{n} \left(\frac{2}{f} - 1\right)^{|U_k - B_k^1| - |U_k - B_k^2|}.
\end{aligned}$$

For any f, $0 \leq f \leq 1$ the number $\left(\frac{2}{f} - 1\right)$ is greater or equal to 1. We are then left to find three n-length Boolean vectors B^1, B^2, and U that maximize $|U - B^1| - |U - B^2|$, *i.e.*, that maximizes $|U - B^1|$ whilst minimizing $|U - B^2|$

Without loss of generality, we can consider that $B^1 = (1, \ldots, 1, 0, \ldots, 0)$, *i.e.* whose first h bits are set with 1. The vector U that maximizes $|U_k - B_k^1|$ is the

reverse of B^1, *i.e.* $U = (U_1, \ldots, U_h, U_{h+1}, \ldots, U_n) = (0, \ldots, 0, 1, \ldots, 1)$, which contains $n - h$ bits set with 1. The n-length Boolean vector B^2 that minimizes $|U - B^2|$ has to set its h bits equal to 1 on the same indices than the ones of U. It is possible if $h \leq n - h$, *i.e.* $h \leq \frac{n}{2}$, which is the case in practice. In other words $B^2 = (B^2_1, \ldots, B^2_h, B^2_{h+1}, \ldots, B^2_{2h}, B^2_{2h+1}, \ldots, B^2_n) = (0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)$. We thus have:

- for k, $1 \leq k \leq h$, $|U_k - B^1_k| - |U_k - B^2_k| = 1 - 0$;
- for k, $h + 1 \leq k \leq 2h$, $|U_k - B^1_k| - |U_k - B^2_k| = 1 - 0$;
- for k, $2h \leq k \leq n$, $|U_k - B^1_k| - |U_k - B^2_k| = 1 - 1$;

Therefore

$$\max_{U, B^1, B^2 \in \mathbb{B}^n} \prod_{k=1}^{n} \left(\frac{2}{f} - 1\right)^{|U_k - B^1_k| - |U_k - B^2_k|} \leq \left(\frac{2}{f} - 1\right)^{2h}$$

and the proof is established.