Anonymously forecasting the number and nature of firefighting operations

Jean-François Couchot[1], Christophe Guyeux[1], and Guillaume Royer[2]

[1]Femto-ST Institute, UMR 6174 CNRS, University of Bourgogne-Franche-Comte, France.

[2]Service Départemental d'Incendie et de Secours du Doubs (SDIS 25), France

May 6, 2019

Abstract

Predicting the number and the type of operations by civil protection services is essential, both to optimize on-call firefighters in size and competence, to pre-position material and human resources... To accomplish this task, it is required to possess skills in artificial intelligence,which are not usually found in a medium-sized fire department. However, such a request may be mandated, for example from specialized companies or research laboratories. This mandate requires the transmission of potentially sensitive information relating to interventions which is not intended to be publicly available. The purpose of this article is to show that a machine learning tool can be deployed and provide accurate results, using a learning process based on anonymized data. Learning on real but anonymized data will be performed using extreme gradient boosting, and the performance of each anonymization will be compared on the number and of interventions per day, and their type.

1 Introduction

For various economic and societal reasons, such as the aging of the population, the closure of small rural hospitals, or the disengagement of the private sector (ambulance drivers) for acts that are not economically interesting, French fire brigades are facing a constant increase in the number of interventions. However, due to the economic crisis and the state debt, the resources allocated to public services in general, and to the fire brigade in particular, are not increasing on their side. The latter must therefore find original solutions to meet growing demand in constant

resource. One solution for the future is to optimize the use of their human and material resources, by pre-positioning vehicles and adapting the size of the guards according to the number, type and location of intervention that an artificial intelligence algorithm could predict.

This solution requires, on the one hand, a database of past interventions that is sufficiently rich and consistent, and on the other hand, know-how in a constantly evolving scientific discipline. This knowledge base is naturally present within the departmental fire and rescue service (SDIS), which collects, for legal and statistical purposes, many data related to each of their interventions. This database contains information on the dates, places and types of interventions, as well as on the interveners and victims. However, if the SDIS has this basis of knowledge useful in the learning phase of an artificial intelligence algorithm, it has neither the know-how nor the human resources to implement such an algorithm.

Indeed, such a realization implies the recovery of explanatory variables by scripts automatically retrieving internet information on past meteorology, ephemerides, epidemiological data, etc. Selecting models from among the various machine learning methods based on decision trees or artificial neurons, as well as feature selection to reduce model complexity, requires time and up-to-date knowledge of machine and deep learning techniques. Similarly, finding good values for algorithm hyperparameters, or proposing resource optimizations based on predictions made, requires the work of computer researchers specialists in artificial intelligence, high performance computing, and optimization.

If the basis of knowledge, with the personal data it contains, is legally protected as long as it remains within the SDIS, its complete transmission to another institution, even if it remains public, is problematic, at least legally. Therefore, the data must be de-identified and then processed by academics, with no intention of public disclosure. However, if anonymization of the data is mandatory to allow such transmission from SDIS to the university, this anonymization should not make the data unusable for any type of prediction. In other words, a fair compromise should be found between the protection of private information contained in the database and the amount of preserved information useful for machine learning algorithms. In fact, the question of whether such a compromise exists and can be found is worth asking.

The objective of this article is to present a concrete case of fine optimization by state-of-the-art techniques, making it possible to guarantee both a sufficiently high privacy given the context (private exchange between fire brigades and academics), while allowing better predictions than what could be obtained with traditional statistical tools. It is therefore a proof of concept on a concrete case study from the SDIS 25 (firemen from Doubs department in France), showing that

a fair compromise is possible, allowing a future optimization of firefighters' resources without paying for it by potential leaks in privacy.

The rest of this article is structured as follows. The case study is presented in the next section, which contains a description of the data under consideration. Section 3 focuses on the problem of de-identification with an overview of most important methods that have been applied on this case study. The database that has been anonymized is then used to learn and predict firemen interventions in Section 4. This article ends by a conclusion section, in which the contribution is summarized and intended future work is outlined.

2 Data Presentation

The data we have to conduct the forecasts are classified by year between 2012 and 2017. Each intervention of the fire fighters of the fire brigade of the Doubs department (a French county of 500,000 inhabitants) is recorded in a file in the form of a line. The attributes of this file are shown in the Table 1 and described as follows:

ID	Station	Reason	Commune	SDate
0	Belfort South	Malaise	Belfort	2018/01/31 08:35

Age	Gender	SAD	Type	Destination
45	Male	No CRA	Other	Belfort Hospital

Doctor	Condition	Location
No	Severe Injury	(47.616, 6.857)

Table 1: Attributes of fire brigade operations data

- *ID* is the ID intervention, which is used in supplementary files;
- *Station* is the fire station name;
- *Reason* is the initial reason for the firefighters' intervention;
- *Commune* is the name of the municipality where the operation took place;
- *SDate* is the starting date of the intervention
- *Age Gender* and *Type* is the age, the gender of the victim, and whether it is a fireman or not;

- *SAD* indicates whether a Semi-Automatic Defibrillator has been used;
- *Destination* gives the subsequent destination, *i.e.* the place where the firefighters transported the victim later;
- *Doctor* specifies whether a doctor was present at the victim's location;
- *Condition* states the victim's condition at the end of the operation;
- *Location* gives the precise location (latitude, longitude) of the intervention.

The Table 2 gives the number of interventions by firefighters per year. As can be seen in this table and as stated in the introduction, the number of firefighters' operations is constantly increasing.

Year	Number of operations
2012	22,960
2013	24,562
2014	26,026
2015	27,750
2016	28,880
2017	31,715
Total	161,813

Table 2: Number of interventions by firefighters per year

3 De-identification problems

This section shows how the fire brigade data were de-identified in order to first predict the number of interventions (Sec. 3.1), then to give the kind of intervention (Sec.3.2).

3.1 Number of interventions per fire station by time slot

The objective of this first part is to have firefighters in each center always in adequacy with the interventions to be carried out by the center's personnel. To know the number of firefighters present and/or available in each center, it is necessary to have an idea of the number of interventions per year, per month, per week, per day, per 3-hour block in each of these centers. The objective is to publish data in the form of tuples (*SDate*, *Station*, *#Operations*) where *SDate* is

the time interval (with variable amplitude as discussed before), *Station* is the station name, or a generalization. Finally, *#Operations* represents the number of actions performed by the fire fighters of the *Station* unit(s) during the time interval *SDate*.

In small rural centers, where the number of interventions is naturally low, it can happen that the hourly amplitude of the study is too low compared to the number of interventions carried out by the fire brigade of the latter and therefore not significant enough to be generalized. It is therefore natural to think of grouping these stations together at the level of the urban community to obtain events that are sufficiently representative in number.

The first question here is: is the number of interventions a sensitive attribute? Clearly yes. This gives importance to a fact. The movement of the Fire Brigade would not take place if the situation had not been critical. For example, if it is known, on the one hand, that a person was sick in a small village and equipped with a centre and that this centre performed an intervention during this period (whereas it almost never does), then there is a high probability that the fire brigade intervened for this person and therefore that the illness worsened.

Two anonymization approaches are used here as direct applications of existing methods. The first is k-anonymity [8] and the second is differential confidentiality [2].

3.1.1 A k-anonymous de-identified dataset.

In order for the article to be self-sufficient, we recall here the definition of the k-anonymity requirement.

Definition 1 (k-anonymity requirement[8]) *Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k individuals.*

In other words, for a given dataset with at least k equivalent records, the probability of re-identifying an individual, for any known given attack A is less than $1/k$.

Thus, only triplets (*SDate*, *Station*, *#Operations*) such as $\#Operations \geq k$ will be provided for further analysis. The others, (*i.e.*, when $\#Operations < k$) will not be used in the further prediction step (since they are removed from the dataset), reducing thus the approach accuracy. This raises the question of choosing a value for the k parameter: a high value decreases the overall probability of re-identification but results in a loss in the data's accuracy. The chosen value k has to ensure an acceptable risk of re-identification for any kind of attack A, *i.e.*, $P(\textit{re-identification}|A) \leq \frac{1}{k-1}$ is lower than a given value.

$$P(\textit{re-identification}|A) = \frac{P(\textit{re-identification}, A)}{P(A)} \quad (1)$$

has to be evaluated for each kind of attack A, namely deliberate attempt at re-identification, acquaintance (*i.e*, inadvertent attempt), or breach. For each kind of attack A, the probability $P(\textit{re-identification}, A)$ must be lower than a commonly acceptable threshold T. Quoting [5], since the dataset will be distributed to researchers only, the average risk threshold T is set to 0.1.

In our context, researchers belong to an academic institution with a confidentiality data agreement, without any particular intent to re-identify records. It is recognized in such a case that $P(\textit{Deliberate Attempt}) \leq 0.4$. The third attack, (breach) can take place if the university loses the dataset. According to [4], it results that $Pr(\textit{Breach}) = 0.27$. We are then left to evaluate $Pr(\textit{Acquaintance})$.

The whole dataset is composed of less that 162,000 operations in the Doubs department (composed of 500,000 inhabitants) which may concern the same individual. The probability of an individual not to be in this dataset is about 1-162000/500000=0.676. Since the average estimated number of well-known contacts is 150, the probability that none of them are in the dataset is approximately equal to 0.676^{150}, which is very close to 0. In this context, the probability of acquaintance is thus equal to 1, *i.e.*, $Pr(\textit{Acquaintance}) = 1$.

The higher the value of $P(A)$, the smaller $P(\textit{re-identification}|A)$ and the more de-identification is required on the data set. One thus have to ensure that $P(\textit{re-identification}|A) \leq \frac{0.1}{1}$, *i.e.* $k = 11$.

Attribute	Generalization Hierarchy
SDate	date-hh:mm:ss → 3-hours block → day → week → month
Station	Station Name → urban community → county

Table 3: Generalization hierarchy for number of interventions per fire station by time slot

A generalization approach can be applied on both attributes *SDate* and *Station* and is represented on Table 3. It is a list of simplifications which can be applied to attribute values, ordered from the smallest intervals to most general ones. Counting the number of operations in an urban community rather than a fire station aims to reduce the number of deletion in the data set to allow for better learning and prediction: there are fewer cases in an urban community than in a fire station where the number of operations per time interval will be less than k. Of course, the results of the predictions will be given at the level of these communities, but many

firefighters live in the metropolitan communities and can move to another fire station if needed. Table 4 gives results of 11-anonymity dataset with respect to the generalization parameters. In each cell, the first number gives the rate of suppressed records whereas the latter is the entropy value [5] expressed as a percentage (compared to the maximum possible entropy for the data set). It can be deduced that the generalization of the starting date to the day, and the fire station to the urban community gives acceptable results both in terms of records loss and entropy.

	SDate	3-hours	**Day**	Week	Month	Year
Fire station	99.8/0.0	99.6/23.1	64.3/70.6	27.6/60.9	5.2/75.4	0.2/99.8
Urban Community	99.8/0.0	96.9/23.1	**32.7/32.8**	7.8/61	0.7/75.5	0.0/99.9
County	99.8/0.19	38.0/38.9	0/42.1	0/61.1	0/75.6	0/100

Table 4: Number of interventions : anonymization by generalization and 11-anonymity

3.1.2 A ε differential private dataset.

Differential privacy [2, 3] is property of anonymization technique that minimizes the privacy impact on individuals whose information is in the database. From a probabilistic point of view, it is not possible for an attacker to identify sensitive data about an individual if his/her information were removed from the dataset. Practically, it may be implemented as noise addition to query results.

Let f be the function that associates to each fire brigade its number of interventions at a given time. If an operation by firemen of this station is deleted, the impact is exactly 1 and the sensitivity of f, usually denoted as Δf, is thus equal to 1. It has been proven that a mechanism that returns $f(x)+y$ where y is the added noise that follows a Laplacian distribution $(0, \frac{\Delta f}{\varepsilon})$ is ε-differential private. A high value of ε leads to small value noise and induces thus a low guarantee of privacy. On the opposite, a small one provides a high probabilistic guarantee against attacks. We are then left to assign a value for the ε parameter with the goal of hiding any individual's presence in the dataset.

According to [6], the value of ε should be bounded by

$$\varepsilon \leq \frac{\Delta f}{\Delta v} \ln \frac{(n-1)\rho}{1-\rho}, \tag{2}$$

where n is the number of lines of our dataset (*i.e.*, at least $n = 22,960$), Δv is the longest distance between two datasets where a line has been removed each time (*i.e.*, $\Delta v = 2$), and ρ is the probability of being identified as present in the database. To be coherent with Sec. 3.1.1, ρ

is set with 0.1. In such a case, ε should be lower than 3.92. The value $\varepsilon = 1$ has been retained here.

3.2 Nature of firefighters' interventions by time slot

To optimize the material and human resources present in each station or urban community, it would be interesting to predict the types of interventions by time interval (year/month/week/ week/day/3-hour block) in each area of interest (station, urban community, department).

For a particular time block of a given amplitude, the types of tasks executed (the reason attribute) are extracted from the data set. The cardinal of this set (in which the equal types are deleted) is naturally lower than the number of interventions found in the Section 3.1. The nature of these interventions is clearly a sensitive data.

There are ≈400 different reasons in the database for firefighters to be involved, some of them overlapping or are very similar to each other. During each intervention, the reason for departure is indicated at the beginning of the intervention, *i.e.* often in an emergency context, leading to a certain number of errors. The finer the granularity, the more errors there are in an emergency situation. To improve data quality, the reasons for firefighters to leave are therefore regrouped into 7 classes that ares *personal assistance*, *road rescue*, *another accident*, *fire or explosion*, *various operations*, *preventive operations*, *other reasons*. This is like applying a low-frequency filter. Once again, it is a question of finding the right compromise between the usefulness of the data and their quality. In all of the following, we only considered data resulting from the grouping in accordance with this filter.

3.2.1 Recursive (c, l)-diversity.

Publishing the types of interventions is critical because if they are not varied enough, then this information can be misused and led to a positive or a negative disclosure. For example, if all the outings that took place on a given date involved heart ailments and if we know that a person was rescued by firemen on that day, we deduce that they had a heart attack. This is the problem identified by Machanavajjhala et al. and named l-diversity [7].

Intuitively, a group of records (bloc, equivalent class) is said to be l-diverse if there are at least l "well-represented" values for the sensitive attributes (which may be a single sensitive attribute, a pair of sensitive attributes, ...). The dataset is said to be l-diverse if each group of records is l-diverse. The notion of "well-represented" is intentionally ambiguous. The fact that l separates values is not sufficient for this definition. A potential refinement could be that the

current values are distributed according to a law approaching uniform distribution. We then find the notion of Entropy l-diversity. However, this constraint is often overly restrictive.

We prefer to take a less restrictive refinement that stipulates that the ratio between the most represented value and the sum of the least $m - l + 1$ represented ones is less than a constant c provided by the user. This definition is known as recursive (c, l)-diversity [7] and is recalled here.

Definition 2 (Recursive (c, l)-Diversity) *In a given q^*-block, let r_i denote the number of times the i^{th} most frequent sensitive value appears in that q^*-block. Given a constant c, the q^*-block satisfies recursive (c, l)-diversity if $r_1 < c(r_l + r_{l+1} + \cdots + r_m)$. A table T satisfies recursive (c, l)-diversity if every q^*-block satisfies recursive (c, l)-diversity.*

The higher the c number is, the more frequently this property is established.

From the experiments in Section 3.1.1 concerning 11-anonymity, we focused on the generalization of fire stations at the level of the agglomeration community and the date of intervention at the level of the day. For this given generalization, we varied l and c both in $\{2, 3, 4, 5, 6\}$. Results are summarized in Table 5 where for each pair (l, c), the rate of suppressed records is given.

In this table and not surprisingly, the number of deleted records is decreasing with respect to c, but increasing w.r.t. of l. Results of de-identification satisfying recursive $(5, 2)$-diversity has been thus retained because it is the only one that does not delete all the data.

l/c	2	3	4	**5**	6
2	93.5	85.0	76.0	**68.0**	61.3
3	99.9	99.7	99.5	99.2	98.6
4	100	100	100	100	100
5	100	100	100	100	100

Table 5: Reasons of interventions: rate of suppressed records with anonymization by generalization, 11-anonymity and recursive (c, l)-diversity

.

3.2.2 Differentially Private Histogram of Operations.

In [9], Xu et al. show how to publish a differentially private compliant histogram which outputs the distribution of a random variable, such as the number of operations with respect ot the

attribute *Reason* of intervention.

The approach is twofold. In the former, for a given time slot (a whole day, e.g.), a histogram is constructed representing the number of interventions performed during this period and whose values are grouped according to the reasons for the intervention of the fire brigade. In the latter, a noise is added with unit-length bins, using the Laplace Mechanism. The resulting histogram is thus published for analyzes. The clustering of reasons for departures into 7 classes (as presented in the beginning of this section) was particularly guided by this step. Indeed, without this grouping, a histogram with potentially 400 bars can be constructed (there are approximately 400 different reasons of intervention, as presented at the beginning of this section). Even with the addition of very low random noise, some of the reasons may appear when they are not at all correlated with an event. By grouping the reasons into 7 classes, the granularity is certainly less, but the added noise is still meaningful.

As in Sec. 3.1.2, ε should be chosen to respect privacy concern. Even if the request executed here on the database is different than the one given in this section, all the variable values of Equation (2) are the same leading to a bound for ε which is 3.92. In what follows, ε is thus set again with $\varepsilon = 1$.

4 Machine Learning Predictions

4.1 General presentation

The objective of this section is now to evaluate whether it is possible to make predictions about the activity of firefighters from anonymized files. We will focus on the number of interventions per unit of time, the type of intervention, and the solicitation per centre. In each case, predictions based on anonymized data will be compared to those based on raw data. More specifically, we will look at whether, from the anonymized data of 2013-2017, we can find out what happened in 2012, as described in the anonymized file of 2012. This predictive ability will be compared to the score obtained by predicting the year 2012 (not anonymized) from the learning on the raw data for 2013-2017. Finally, the 2012 prediction based on the anonymized 2013-2017 data will be compared to the de-anonymized 2012. Note that we have chosen to predict 2012 from 2013-2017, and not 2017 from 2012-2016, because the year 2017 saw its number of interventions explode due to a disengagement of the private sector (ambulance drivers) artificially inflating firefighters' interventions, and this for reasons that are difficult to predict because they are no longer linked to human activity: instead of predicting the future, we are reconstructing a potentially unstored

past.

In order to achieve this supervised learning, we had to recover a collection of explanatory variables that could potentially explain the number, type and location of interventions. We have assumed that these interventions are directly related to human activity (for example, there is less intervention at night, because people sleep), which itself changes according to the time of year (holidays, seasons...), the weather, etc. These explanatory variables, for each hour of the period under consideration, are publicly available on the Internet, and have enabled us to recover with some precision the 2012 interventions from those of 2013-2017.

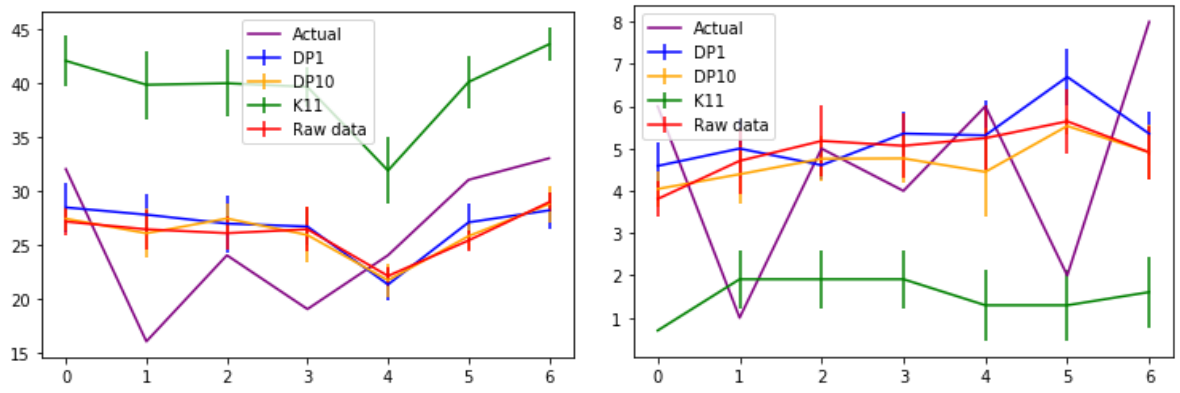
In detail, the following numerical variables were recovered from the MétéoFrance site for the three weather stations closest to the Doubs (Nancy, Dijon and Basel): wind direction, humidity, dew point, precipitation during the previous hour, and during the last 3 hours; pressure, and its variation lods over the last 3 hours, temperature, wind speed, and finally visibility. At the calendar level, we have added the year, month, day in the week (Monday...), in the month (1.2, ..., 31) and in the year, in order to identify days different from the normal (national holiday, Christmas...). Epidemiological data have also been added on the incidence of influenza, chickenpox and diarrhoea over the past week, collected from the Sentinel network. Finally, since the Doubs department is rich in mountains, forests and rivers, in a temperate region, we occasionally have heavy rainfall leading to sudden variations in the height of the rivers. The latter lead to floods, requiring assistance to people. Also, the heights of six rivers have been added, with their variations over the past hour.

In the following, we will present the prediction results from approaches that can be obtained using the explanatory variables on original data, and finally what is found using the anonymized version of the data. Finally, it should be noted that the machine learning algorithm used here is the extreme gradient boosting (XGBoost), with the default values as hyperparameters [1]. For each set of prediction attempts, 5 experiments have been done with distinct seeds for random initialization. Each curve represents the curve of the means and the standard deviation is always displayed with vertical bars.

4.2 Predicting the number of interventions

In this section, the objective is to predict the number of interventions per fire station by time slot. The de-identification has shown that an acceptable trade-of between the number of suppressed data, the entropy and the duration of the study is obtained by merging data inside Urban Communities and for a duration equal to the day (see 4).

Let us first recall that ensuring 11-anonymity had a cost: a number of lines have been deleted. More precisely, 32.7% of the interventions were removed, and therefore a factor multiplying the number of predicted interventions by 1/(1-32.7/100) = 1.486 will be considered. For this method, this corresponds to an adjustment achieved by a systematic increase in forecasts of about 49%.



(a) Grand Beançon, (b) Grand Pontarlier,

Figure 1: Predictions Week # 2, 2012

Three anonymisation methods have been applied, namely 11-anonymity and Differential Privacy with $\varepsilon = 1$ and $\varepsilon = 10$. Figure 1 and all subsequent ones compare the results of predictions from the original data, anonymized data using these three methods, and the mean (which is the simplest prediction). Each time, these predictions are given for the week between January 8 and 14, chosen as an arbitrary example. All these results focus on two agglomeration communities, namely Grand Besançon and Grand Pontarlier. These two agglomerations were chosen because their demographic characteristics are significantly diverse. The former has about 200,000 inhabitants living in 68 municipalities, mainly urban over a surface area of 528.6 km2, representing approximately 379 inhabitants/km2. The latter is composed of approximately 27,000 inhabitants who live in 10 municipalities, mainly rural, over an area of 154 km2, that is, approximately 175 inhabitants/km2.

In this figure and in all the following, Actual (in purple) always denotes what really happens during this week. Concerning predictions, K11 (in green) is the average curve of predictions through 11-anonymity concept. DP1 and DP10 are the mean curves after Differential Privacy based anonymization with $\varepsilon = 1$ and $\varepsilon = 10$ respectively. Red curves represent data forecast from original row data (*i.e.*, non de-identified data).

Let us explain results obtained for the agglomeration of Grand Besançon (Figure 1(a)). First of all, the average number of interventions for this agglomeration is 28.1 with a standard deviation of 7.7. The Mean Absolute Error (MAE) when considering this average as a prediction

of reality is 6.6. Any prediction based on intelligence must reduce this error.

Note that all predictions based on Differential privacy and on row data are consistent: the standard deviation for each prediction is about 6.6. Here, the prediction with data anonymized with 11-anonymity is over-estimated: as already announced, this method has lead to a suppression of 32.7% of data indeed. But only a few of removed data concerns Grand Besançon and the forecast accuracy for this urban community was thus sufficient enough. However all the predictions with this generalization based anonymization method have been increased by 49% leading here to a over-estimation. In this agglomeration, even if the predictions are not extremely accurate, we can see that they follow the same trend as reality: a relative decreasing until the middle of the week with a more or less rapid ascendancy thereafter. The mean average errors w.r.t the chosen anonymisation method are reported in Table 6. It can be seen in the latter that the predictions on data anonymized by Differential Privacy have the same level of accuracy as those from the original data.

The results are much less homogeneous for the Pontarlier urban community (Figure 1(b)). In this case, the predictions from the data anonymized by 11-anonymity are far below reality and other predictions. This is explained by the fact that in this urban community, the average number of interventions for this week is 4.7 with a standard deviation of 2.6. Many data concerning this agglomeration community are thus deleted by the 11-anonymity method. The average number of interventions using this latter anonymization method is indeed 1.7, after the adjustment of the data by 1.5. However, this result is far below reality. The other approaches based on Differential Privacy anonymization methods give forecast which are in the consistent order of magnitude. Regarding the MAE of predictions (Table 6) and as in the other agglomeration community, predictions are as accurate when using data anonymized by Differential Privacy as when embedding raw data.

Another positive point is that we find, in general, the same relative importance of each explanatory variable: the same causes explaining the number of interventions are highlighted (causality is not confused): the five most important features as provided by the plot_importance function (namely, the year, wind direction, day in the year, humidity, and water level of the Doubs River) are the same, but not in the same order. Let us also note to relativize that, on anonymized data, we obtain predictions that are not totally meaningless (compared to the average), while:

- no model selection (choice of the machine learning algorithm) has been performed;

	Grand Beançon	Grand Pontarlier
Average number of intervention	28.1	4.7
Mean	6.6	2.0
11-anonymity	16.0	3.6
DP ($\varepsilon = 1$)	5.6	1.9
DP ($\varepsilon = 10$)	5.5	1.9
Raw data	5.7	1.9

Table 6: Mean Average Error with respect to anonymization method

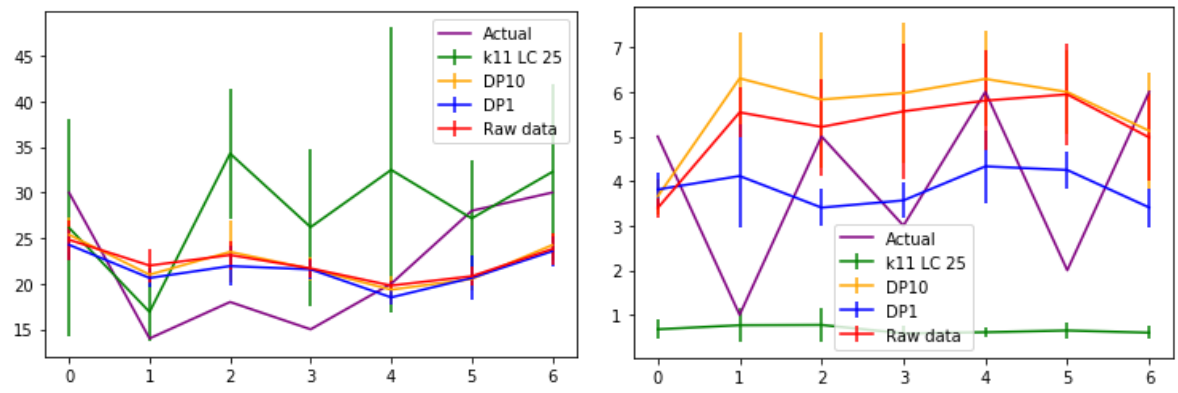
- no preliminary step was taken to select explanatory variables;
- no attempt was made to optimize the many hyperparameters of the XGBoost.

4.3 Predicting the nature of interventions

In this set of experiments, two anonymisation methods have been applied. The former is 11-anonymity combined with recursive $(5, 2)$-diversity and the latter is histogram of operations compliant with Differential Privacy (with $\varepsilon = 1$ and $\varepsilon = 10$). For the same reasons as above, this study focuses on the two agglomeration communities, Grand Besançon and Grand Pontarlier. This article focuses only on two types of intervention, namely *personal assistance* and *road rescue*. Personal assistance is indeed very frequent and can usually be managed by several services: the SAMU, private ambulances and fire brigades. In contrast, road accidents are more infrequent (and predictable with probably less accuracy), but are systematically handled by firefighters. Results of predictions are shown in Figures 2 and 3. The former deals with personal assistance whereas the latter focuses on road accidents. As in the previous section and for the same reasons, this figures focus on two agglomeration communities, namely Grand Besançon and Grand Pontarlier and the color codes are the same than in previous section.

Let us first focus on personal assistance. As in the previous section, the number of interventions realized for this reason of departure is overestimated when anonymization is achieved by 11-anonymity and recursive (5,2)-diversity when the urban community is Grand Besançon and underestimated otherwise. It happens that data containing this reason may be deleted by this method.

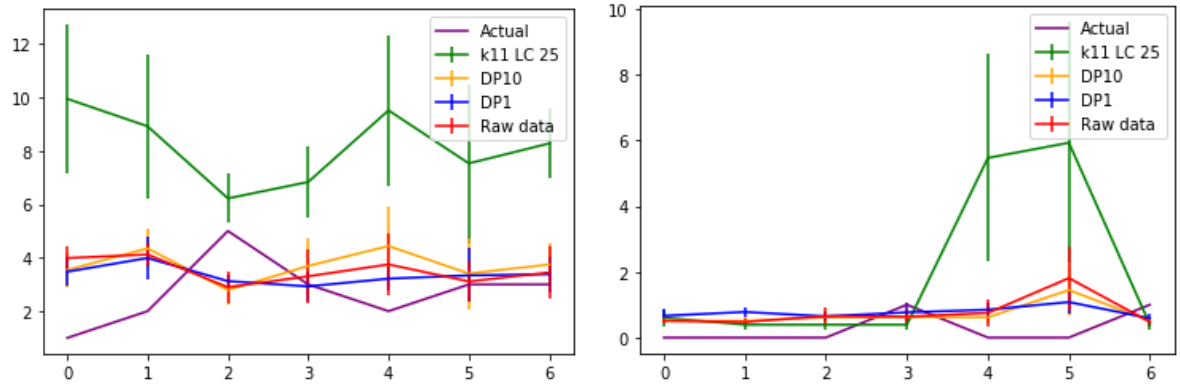
For a medium-sized urban community such as Besançon, the trend is observed also on



(a) Grand Beançon, (b) Grand Pontarlier,

Figure 2: Personal assistance, Predictions Week # 2, 2012

anonymized data, even if it is slightly overestimated. This is explained by the fact that the number of interventions for this reason of intervention has increased steadily over the years (between 2012 and 2017) and that 2012 is therefore the year in which these exits have been the least numerous. For the small urban community of Pontarlier, the trend is also found even on data anonymized by the method combining histograms and differential confidentiality. For ε equal to 1 (which guarantees acceptable safety), predictions close to the mean are found.



(a) Grand Beançon, (b) Grand Pontarlier,

Figure 3: Road intervention, Predictions Week # 2, 2012

Road accidents are quite uncommon and therefore more difficult to predict, especially in small rural communities. In this context, it seems even less relevant to apply the 11-anonymity and recursive (5,2)-diversity based anonymization method to make subsequent predictions. This is confirmed by the curves of the figures 3(b) and 3(a). The standard deviation for this anonymization method are indeed very large, and forecast a very far from reality.

The noise added by the method combining histograms and differential confidentiality is

sufficiently limited even when ε=1. Indeed, the general trend is found with almost as much precision on data anonymized by such an approach as on raw data, *i.e.*, on non-anonymized data. This trend is numerically validated by the values given in Table 7, which summarizes the mean absolute errors by agglomeration community, by type of intervention and according to the chosen anonymization method. As in the previous table on prediction errors concerning the number of interventions, it can be seen here that the histogram method with differential privacy allows to obtain predictions as precise as those obtained on raw data.

	Grand Beançon		Grand Pontarlier	
	Personal assistance	Road accident	Personal assistance	Road accident
Average number of intervention	24.1	3.4	4.0	0.6
Mean	5.6	2.0	1.7	0.8
11-anon. + recursive (5,2) diversity	7.6	3.9	3.3	0.9
DP ($\varepsilon = 1$)	4.5	2.0	1.7	0.8
DP ($\varepsilon = 10$)	4.6	2.0	1.7	0.9
Raw data	4.5	2.0	1.6	0.8

Table 7: Mean Average Error with respect to anonymization method

5 Conclusion

"'Can we predict and with which accuracy the number (1) and nature (2) of firefighters' interventions in a geographical area while respecting the privacy of the victims they rescued?" This article is a positive answer. In both the quantitative (question (1)) and qualitative (question (2)) domains, this article shows that differential confidentiality based approaches provide more accurate results than generalization and suppression ones. It is possible to use privacy-respecting (*i.e.*, properly anonymized) data to guess an accurate behavior.

It should be noted that the variable ϵ was deliberately set to 1 to ensure a high level of privacy. By increasing this value (up to the calculated threshold 3.9), the obtained results would have been even more accurate.

The prospects for this work are numerous. We will first study the possibility of predicting

the places of intervention, knowing that this attribute is very critical, because it almost allows the victim to be identified.

This study has been supported by the EIPHI Graduate School (contract "ANR-17-EURE-0002"), by the Interreg RESponSE project, and by the SDIS25 firemen brigade.

References

[1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[2] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[3] 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.

[4] Khaled El Emam and Luk Arbuckle. *Anonymizing health data: case studies and methods to get you started.* " O'Reilly Media, Inc.", 2013.

[5] Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, et al. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16(5):670–682, 2009.

[6] Jaewoo Lee and Chris Clifton. How much is enough? choosing ϵ for differential privacy. In Xuejia Lai, Jianying Zhou, and Hui Li, editors, *Information Security, 14th International Conference, ISC 2011, Xi'an, China, October 26-29, 2011. Proceedings*, volume 7001 of *Lecture Notes in Computer Science*, pages 325–340. Springer, 2011.

[7] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 24–24. IEEE, 2006.

[8] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, technical report, SRI International, 1998.

[9] Jia Xu, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, Ge Yu, and Marianne Winslett. Differentially private histogram publication. *The VLDB Journal—The International Journal on Very Large Data Bases*, 22(6):797–822, 2013.