UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

UNIVERSITÉ DE FRANCHE-COMTÉ

THÈSE DE DOCTORAT

DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PRÉPARÉE À L'UNIVERSITÉ DE TECHNOLOGIE DE BELFORT-MONTBÉLIARD

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Informatique

par

ROXANE ELIAS MALLOUHY

Predictive analysis of time series in various application contexts

Analyse de flux par des techniques de séries temporelles

Thèse présentée et soutenue à Belfort, le 5 JANVIER 2023

Composition du Jury :

PR GUYEUX CHRISTOPHE	Université Bourgogne Franche-Comté	Directeur de thèse
PR ABOU JAOUDE CHADY	Université Antonine	Co-directeur de thèse
PR DEMERJIAN JACQUES	Université Libanaise	Rapporteur
PR BRUNIE LIONEL	Insa Lyon	Rapporteur
PR MAKHOUL ABDALLAH	Université Bourgogne Franche-Comté	Examinateur
PR HIJAZI ABBAS	Université Libanaise	Examinateur

ABSTRACT

Predictive analysis of time series in various application contexts

Roxane Elias Mallouhy
University of Bourgogne Franche Comté, 2022

Supervisors: Christophe Guyeux and Chady Abou Jaoude

Emergency medical transport in France is triggered by the dispatch of an ambulance, either by the SMUR (SAMU), by a private ambulance company, or by the fire department after dialing one of the emergency numbers (15, 18, or 112). Since accidents are related to human activities, which in turn depend on the time of day, season, weather, climate, special events, etc., the emergency response (ER) is not hazardous but predictable. For example, falls on icy surfaces occur in winter, while drowning in swimming pools occurs in summer. In addition, accidents are more likely to occur on a holiday eves rather than a normal day when people are resting and sleeping at home. Thus, the flows of these actors are predictable to some extent, especially because of their seasonality. Being able to predict such operations makes it possible to put in place strategies for planning that could be very helpful in managing the emergency sector, which is currently in crisis. For example, predicting the need for volunteer or non-volunteer firefighters allows the department to anticipate firemen resources to match demand. Forecasting firefighter interventions for the short term (two or three months) allows for better planning for paramedic leave at the emergency level, while forecasting for the long term (several years) facilitates planning of future human and material resources.

In this context, the collection of data from different streams varies (hospitals of Belfort-Montbéliard, Nord-Franche-Comté, Doubs region, Île-de-France) over periods ranging from a few years to twenty years, the aim being to use these different flows both to analyze their dynamics and to make more or less long-term forecasts. Some of these data streams have already been used in a supervised learning approach that requires the continuous collection of a set of explanatory variables (related to meteorological, ephemeral, epidemiological data, etc.), which proves to be complex for an operational device: scripts

need to be set up to retrieve these variables on an hourly basis, scheduled periodically for new machine learning, etc. As a result, different approaches have been studied and applied to different firefighter datasets provided by the fire and rescue department, with the main objective being to study such operations for better future planning and management of the emergency response at lower complexity.

Firstly, statistical analysis models (AR, MA, ARIMA) are compared with supervised learning approaches, and newer techniques such as Prophet are observed. Each algorithm is explained with best fit parameters and statistical features are calculated and then compared between applied models on the same dataset.

Second, exponential smoothing (Simple, Holt, and Holt-Winters) is applied to examine and verify the principal components and trends in the firefighter dataset, showing that the number of firefighters' interventions is not a simple random process, but is influenced by hourly and weekly changes related to human activities. Such a technique is a prominent tool to use in such a study that provides reliable forecasting where statistical characteristics and graphical exploration of data have been presented to find the best Exponential Smoothing technique.

Then, the number of firefighter incidents in the COVID-19 period was examined to verify the efficiency of the predictive model in such critical and rare events. The first step was to select the most relevant features of the dataset, then the breakpoint was detected and last anomalies were replaced. In each step, the prediction of firemen interventions was performed using the XGboost algorithm.

Afterwards, the extension of the study, taking into account not only the number of fire calls but also the type of calls, leads to a better understanding of the structure of the emergency departments in the studied regions of France by generating 14 sub-datasets of different types of missions.

Finally, the clustering of the different departments of firefighters in Île-de-France according to similarity is suitable to analyze the activities of the fire department. The aspect of forecasting is not only considered but also analyzed in depth to gain knowledge and efficient outcomes in this area.

KEYWORDS: Time series Forecasting, Clustering, Data mining, Machine learning, Firefighters' interventions, Feature selection, Breakpoint, Anomalies detection, Statistical Features.

RÉSUMÉ

Analyse de flux par des techniques de s eries temporelles

Roxane Elias Mallouhy
Université de Bourgogne Franche Comté, 2022

Encadrants: Christophe Guyeux et Chady Abou Jaoude

Le transport sanitaire d'urgence est enclenché, en France, suite à l'appel à un des numéros d'urgence (15, 18 ou 112), et suite à cet appel une ambulance est envoyée, provenant soit du SMUR (Structures Mobiles d'Urgence et de Réanimation) ou le SAMU (Service d'Aide Médicale Urgente), soit d'une entreprise d'ambulanciers privés, soit des sapeurs-pompiers. Les accidents étant liés à l'activité humaine, qui elle-même est conditionnée à l'heure dans le jour, à la saison, au temps qu'il fait, etc., la sollicitation pour du secours à personnes n'est donc pas aléatoire, mais prévisible. Ainsi, les chutes sur des plaques de verglas se produisent en hiver, quand les noyades en piscine privée d'extérieur se produisent quand il fait bon. Les flux de ces trois opérateurs sont donc prévisibles, dans une certaine mesure, notamment du fait de leur caractère saisonnier. Et parvenir à les prévoir rend possible la mise en place de stratégies de planifications, qui pourraient aider grandement à la gestion de ce secteur actuellement en crise. Par exemple, être en mesure de prévoir la sollicitation à l'horizon de quelques heures, chez les pompiers, leur permet d'anticiper le besoin en pompiers volontaires. Avoir une visibilité à court terme (deux ou trois mois) permet de planifier au mieux les congés des ambulanciers ou au niveau des urgences, quand une visibilité à long terme (plusieurs années) aide à la planification des besoins futurs, tant matériel qu'humain.

Dans ce contexte, la collecte de données de différentes filières varie (hôpitaux de Belfort-Montbéliard, Nord-Franche-Comté, Doubs, Île-de-France) sur des périodes s'étalant de quelques à une vingtaine d'années L'objectif consiste à exploiter au mieux ces flux, tant pour en analyser la dynamique que pour être en mesure d'effectuer des prévisions à plus ou moins long terme. Certains de ces flux ont d'ores et déjà été exploités dans une approche d'apprentissage supervisé, qui nécessite la collecte en continu d'un certain

nombre de variables explicatives (liées à la météorologie, aux données d'éphémérides et d'épidémiologie, etc.), ce qui s'avère complexe à mettre en oeuvre pour un dispositif opérationnel : des scripts doivent être mis en place pour récupérer à chaque heure ces variables, planifier périodiquement de nouveaux apprentissages automatiques, etc.

En conséquence, différentes approches ont été appliquées sur différents jeux de données de pompiers fournis par le service départemental d'incendie et de secours (SDIS), avec l'objectif principal d'établir une meilleure planification et une meilleure gestion future des pompiers à moindre complexité.

Tout d'abord, des modèles d'analyse statistique (AR, ARMA, ARIMA) sont comparés à des approches d'apprentissage supervisé, et des techniques plus récentes telles que Prophet sont observées. Chaque algorithme est expliqué avec les meilleurs paramètres, les caractéristiques statistiques sont calculées puis comparées entre les modèles appliqués sur le même jeu de données.

Deuxièmement, des méthodes Exponential Smoothing (simple, Holt et Holt-Winters) sont appliquées pour examiner et vérifier les principales composantes et tendances de l'ensemble de données des pompiers, montrant que le nombre d'interventions des pompiers n'est pas un simple processus aléatoire, mais est influencé par des changements horaires et changements hebdomadaires liés aux activités humaines. Une telle technique est un outil important à utiliser dans une telle étude qui fournit des prévisions fiables où les caractéristiques statistiques.

Ensuite, le nombre d'incidents de pompiers dans la période COVID-19 a été examiné pour vérifier l'efficacité du modèle prédictif dans un tel événement critique et rare. La première étape consistait à sélectionner les caractéristiques les plus pertinentes de l'ensemble de données, puis le breakpoint a été détecté et les dernières anomalies ont été remplacées et à chaque étape, la prédiction des interventions des pompiers a été effectuée par l'application de l'algorithme XGboost.

Par la suite, l'extension de l'étude, prenant en compte non seulement le nombre d'appels au 18 mais aussi le type d'appels, conduit à une meilleure compréhension de la structure des services d'urgence dans les régions de France étudiées, en générant 14 sous-ensembles de données de différents types de missions.

Enfin, le regroupement des différents services de sapeurs-pompiers d'Île-de-France selon la similarité est propice à l'analyse de leurs activités. L'aspect de la prévision est non seulement pris en compte, mais également analysé en profondeur pour acquérir des connaissances commerciales et des résultats efficaces dans ce domaine.

Mots clés : Séries temporelles, Clustering, Data mining, Machine learning, Interventions des sapeurs-pompiers, Sélection des attributs, Breakpoint, Détection d'anomalies, Paramètres statistiques.

ACKNOWLEDGEMENTS

This project would not have come to fruition without the support of many people who assisted me in one way or another from the first phase to the last. I greatly appreciate the insights and advice and express my deepest gratitude to a number of people.

Prof. Christophe Guyeux, my thesis director at UBFC whose sincerity and encouragement I will never forget. Pr. Guyeux was an inspiration on the way to my dissertation and would never have been possible without him. He guided me generously from the beginning and helped me master the subject with a regular follow-up. I am grateful for the extraordinary experience and opportunity to work with him.

Dr. Chady Abou Jaoude, my thesis co-director at University Antonin, for his constant encouragement, guidance and support along the way on this thesis. My sincere appreciation for selecting me for this project.

Prof. Abdallah Makhoul, for his valuable feedback, comments and advices during the experiments and writing of the scientific articles.

My parents, husband and friends, for their constant motivation, strong support and tremendous encouragement to get me to this point despite all the challenges. To my **children** I give everything, including this achievement.

Sense of respect to **Father Roland Hamid Awkar**, who inspired me. Without him, I would not have been able to embark on this journey.

Finally, I would like to thank the **jury members** for examining and reviewing my work.

CONTENTS

List of abbreviations

AI . . . . . Artificial Intelligence

AR . . . . . Auto Regression

ARIMA . . . . . Autoregressive integrated moving average

CART . . . . . Classification And Regression Tree

ED . . . . . Emergency Department

ER . . . . . Emergency Response

FA . . . . . Factor Analysis

LDA . . . . . Linear Discriminant Analysis

LightGBM . . . Light Gradient Boosting Machine

LSTM . . . . . Long short-term memory

MA . . . . . Moving Average

MAE . . . . . Mean Absolute Error

ML . . . . . Machine Learning

PCA . . . . . Principal Component Analysis

PPDM . . . . . Privacy Protected Data Mining

RMSE . . . . . Root Mean Squared Error

RNN . . . . . Reccurent Neural Network

SAMU . . . . . Urgent Medical Aid Service

SDIS . . . . . Departmental Fire and Rescue Service

SMUR . . . . . Mobile Emergency and Resuscitation Service

SVD . . . . . Truncated Singular Value Decomposition

TSF . . . . . Time series forecasting

XGBoost . . . . EXtreme Gradient Boosting

I

DISSERTATION INTRODUCTION

1

INTRODUCTION

Most artificial intelligence (AI) research focuses on solving real-world societal problems such as health services, economics, education, finance, agriculture, commerce, biology, transportation, entertainment, and more. The process begins with integrating a particular piece of information, analyzing the data, obtaining results, and using those insights to improve decision making and life outcomes.

Machine Learning (ML), an application of Artificial Intelligence, is the current revolution that is emerging these days and will affect life in almost every field, mainly because ML offers a potential use for the huge amount of data available in the world.

Data mining, which uses techniques developed by ML is the process of dealing with and analyzing large scattered datasets to extract hidden useful information. It can also be referred to as data or knowledge discovery. It involves the use of statistical methods, artificial intelligence, algorithms, and software to identify patterns and correlations among millions of records to predict future observations and trends for various objectives. Data mining techniques are widely used in organizations to increase the efficiency of operations, such as marketing, retail, banking, fraud detection, cybersecurity, risk management, mathematics, education, and medicine. It is also used in social networking, robotics, signal processing, computer vision, mobile computing, and many more. Data mining is mainly applied to make better or faster business decisions, extract relevant information, or remove redundancy and noise from data.

Broadly speaking, the data mining process begins with the identification of business problems in order to analyze data sources, i.e., databases. Then the data is collected, examined, processed, and transformed as needed. After that, the dataset is tested and evaluated to obtain analysis results.

According to the convention between the University of Bourgogne Franche-Comté (UBFC), Belfort-France and Antonin University (UA), Hadath-Lebanon, the presented work was carried out in alternation between the DISC (Département Informatique et Systèmes Complexes) in France and the ticket laboratory in Lebanon.

Throughout the rest of this paper, the author will be referred to as “we” rather than “I”. The reason for this is that this paper presents research that was conducted in a collaborative environment, as part of teams in both research laboratories.

This chapter contains an introduction to the work being done in this dissertation. It addresses the general context and highlights of this thesis. The contributions of this work are also briefly introduced.

1.1/ Supervised versus unsupervised Machine Learning

In the field of machine learning and artificial intelligence, there are two different approaches: supervised [Cunningham et al., 2008] and unsupervised learning [Celebi et al., 2016]. The main difference is that supervised learning uses labeled data to make predictions as it involves predetermined output, while unsupervised learning does not specify target attributes. Moreover, in supervised learning, the algorithm must iteratively learn from the training dataset to predict the output, while unsupervised learning models automatically discover the organization of the unlabeled data, still, some human intervention is required to validate the output variables.

1.1.1/ Supervised learning

Supervised learning algorithms are further divided into classification and regression algorithms.

- Classification: can be either binary, i.e., there are only two groups for categorizing the output, or multinomial, i.e., the data is classified into three or more groups. Basically, the data inputs are divided into a certain number of classes or categories based on the training dataset. For example: filtering spam emails into different inbox folders [Youn et al., 2007], handwritten letters/numbers detection [Kavallieratos et al., 1997], document classification, web text classification [Saha, 2011], credit card fraud detection [Shen et al., 2007], and many more. Some popular classification algorithms are: Decision Tree [Wu et al., 2008], Naive Bayes [Webb et al., 2010], K-Nearest Neighbor [Peterson, 2009], and Support Vector Machine [Cortes et al., 1995].
- Regression: prediction of numerical values by understanding the relationship between input and output variables based on different data points. For example, it can be used in predicting house prices in the future [Ho et al., 2021], predicting customer purchases [Martínez et al., 2020], targeted advice and marketing [Perlich et al., 2014], predicting stock prices [Emioma et al., 2021], and other

applications. Linear regression [Galton, 1886], logistic regression [Cox, 1958] and polynomial regression [Ostertagová, 2012] are all common types for regression algorithms.

1.1.2/ UNUPERVISED LEARNING

Unsupervised learning automatically discovers the structure of unlabeled data and is mainly used in two major techniques: clustering and dimensional reduction.

- Clustering: unlabeled data are grouped based on their similarities and differences. Clustering can be used in various fields, such as: genetic grouping [Miller et al., 2020], market segmentation [Schlager et al., 2022], medical imaging [Angelini et al., 2022], and others. Many algorithms exist to accomplish clustering techniques: K-mean [Hartigan et al., 1979], DBSCAN [Hahsler et al., 2019], BIRCH [Zhang et al., 1996], and many more.

- Dimensional technique: theoretically used in the preprocessing step when the dimension of the dataset is too large and has an enormous number of attributes. This technique reduces the number of features to a feasible size while maintaining integrity. It also decreases the overall computation time, avoids the problem of overfitting, and removes noise from the data. There are many techniques for reducing the dimensionality of the dataset: Principal Component Analysis (PCA) [Jolliffe et al., 2016], Factor Analysis (FA) [Taherdoost et al., 2014], Linear Discriminant Analysis (LDA) [Balakrishnama et al., 1998], Truncated Singular Value Decomposition (SVD) [Hansen, 1990] and others.

1.2/ USE CASE: THE FIREFIGHTERS' INTERVENTIONS

French firefighters are not only called to put out fires. Their mission goes beyond that. They are well trained and equipped to be the first responders to both medical and domestic emergencies. Due to the aging of the population and the restructuring of hospitals, firefighters are increasingly called to medical-social missions, especially as private paramedics. In fact, 80% of the activity of firefighters today is devoted to missions that are mainly not their responsibility [Gotschaux, 2019]. At the end of 2020, there are about 251900 firefighters who have performed 4290700 missions in fires, emergency human assistance, traffic accidents, and other miscellaneous missions [col, 2020]. Among them 78.25% were volunteers, 16.59% were professionals and 5.16% were military firefighters [pom, 2022].

The French healthcare system is considered one of the best in the world, but in recent years it has experienced an unprecedented crisis in hospitals. Under-staffing of doctors and nurses is due to budget cuts that result in intensive use of existing staff while the number of patients increases. Hospitals have a funding system where funds are allocated based on the number of patients. Hence, some have cut their budgets, and others have ceased operations altogether. As a result, firefighters' interventions appear to have increased to make up for shortages at hospitals.

On the other hand, fire brigades in France have been responsible for transporting patients from the hospital to their homes during this crisis. Their duties were extensive in this regard, as the aging population in France requires additional care and support. Budget cuts in hospitals have made it necessary to get people back home as quickly as possible, as the number of beds is saturated. This was the main task of the firefighters.

All of these and other factors have driven up the number of firefighters' missions: non-emergency calls have increased since 2009. The firefighters are on strike today, demanding from the Ministry of Interior a continuous increase in funding and better working conditions. Above all, they demand action for the public service, which has been neglected by the government [eps, 2019]. Therefore, optimizing the use of their resources as needed will improve the efficiency of the response in terms of the number of personnel and latency during peak periods. This will also directly reduce financial resources. It is important to emphasize that the influx of firefighters is somewhat related to climate, time, and some events. Floods in summer are an event that occurs less frequently than in winter. Or fires are more likely to occur in the summer than in the cold weather. Therefore, predicting such operations could be done using machine learning approaches since fire department operations are directly related to human activities: accidents are more likely to occur during the day than at night or during a vacation.

The data used in this study particularly enables the use of time series forecasting in decision making by estimating future trends and scale. Consequently, the concept of applying machine learning techniques to tailor the need for firemen operations to their demands is doable to organize the firemen human and material resources.

1.3/ Main Contributions

The main contributions in this dissertation fall under the optimization of the firemen resources. These contributions can be summarized as follows:

1. First, three different well-known time series algorithms were implemented: Autoregression (AR), Moving Average (MA), and Autoregressive Integrated Moving Average (ARIMA). The dataset used contains information about firefighters' in-

terventions in the Doubs-France region from 01/01/2006 00:00:00“ to 31/12/2017 23:00:00”. All data was registered by the fire and rescue department SDIS 25 in blocks of one hour. The statistical parameters of these three different machine learning algorithms were calculated to predict the number of firefighting operations, a prophet was implemented, and a comparison between the applied algorithms and related work using the same dataset was performed. The results validated the efficiency of the ARIMA model. On the other hand, we applied another technique to predict time series data, namely exponential smoothing. The purpose behind using this technique is that it takes into account three main components: trend, level, and seasonality, which could be a major component of the fire department dataset, since the number of deployments is somewhat related to weather, climate, and date. Simple exponential smoothing, Holt, and Holt-Winters were applied over the 2015-2019 period in which the dataset was prepared and the model was built to realistically predict the future.

2. In the second contribution, we apply machine learning techniques to tailor the need for firemen operations during the sensitive global pandemic COVID-19. The experiments were applied to a dataset from 2016 to 2021 provided by the same fire and rescue department, resulting in the detection of a breakpoint associated with the epidemic. The first step was to select features by applying the feature importance method, then the breakpoint was detected and finally anomalies were replaced. In each step, the prediction of firemen interventions was performed using the XGboost algorithm. The results show promising accuracy in predicting regular or irregular events such as the COVID-19 epidemic.

3. Alongside with the prediction of fire responses using different techniques in the first two contributions, we present the types of interventions related to 14 different subsets of data for each type of possible category (childbirth, fire, suicide, traffic accident, drowning, public road fire, flood, heating, emergency human aid, help for people, public road accident, brawl, witness, and wasp). All the studies proposed in recent years consider different metrics but never the type of operations. Integrating the deployment category is feasible and shows accurate results, however, the subject of the deployments and the size of the sub-datasets also play an important role in the accuracy.

Afterwards, we merge the two datasets of the number of firemen interventions and the type of missions to investigate the effect of adding explanatory variables to the existing attributes, which are considered to be very simple and refer only to the date and time. The datasets for each category were created by grouping the deployment types and creating 14 sub-datasets for each fire deployment independently. Besides, these datasets were merged with the huge original data used in previous

contributions, and finally, the best features for each sub-dataset were selected using the feature importance technique to reduce computation time and storage requirements. All evaluations were performed using two well-known boosting machine learning algorithms, XGBoost and LightGBM.

4. Finally, we proposed the application of k-mean clustering on a dataset of firemen interventions to group the number of fire brigades into different clusters, leading to a better understanding and interpretation of firefighters' operations in the Île-de-France, particularly in Yvelines. The dataset was preprocessed, the outliers were removed in some areas showing anomalies, and the k-mean clustering technique was performed after choosing the optimal parameter k for the number of clusters. The partitioning of the clusters was analyzed to inspect which criteria were responsible for the partitioning.

1.4/ Dissertation Outline

The rest of this dissertation is organized as follow: Chapter 2 provides a general overview of machine learning and its challenges, and highlights key applications of data science in the emergency department. Chapter 3 hands over an overview of the key features and advances of time series and presents the state of the art using the dataset of firefighters used in this thesis. In addition, Chapter 4 reveals the infrastructure used to perform the techniques of ML. Moreover, in Chapter 5, the first contribution is made by presenting the dataset, implementing different ML techniques, and predicting the number of firefighters in many experiments. On the other hand, in Chapter 6, the COVID-19 period was included in the forecast to train the model for predictions in sensitive periods. In Chapter 7, not only the number of fire departments but also the type of operations are presented using two different approaches. In Chapter 8, a completely new dataset was utilized and the deployments were divided into different clusters. Finally, a conclusion and an outlook are given in chapter 9.

II

ML in Emergency Departments: overview, techniques and tools

This part provides an overview of the applications of ML and data mining in various fields, including healthcare, specifically ER. The first chapter highlights the applications of machine learning in various domains; the second chapter focuses specifically on the use of time series in the ED and discusses recent advances in this area. The third chapter discusses the principles of time series forecasting and their application to the firefighters dataset. The methods used in this dissertation and the differences between their algorithms are then explained, and the final chapter discusses the infrastructure used in this thesis.

2

ML Applications and challenges

The use of machine learning is increasing every day, and new techniques are being developed regularly. ML is used for everyday problems in various fields, such as sentiment analysis, language modelling, text/image classification, object recognition, semantic segmentation, question answering, machine translation, speech recognition, time series analysis, and many others.

2.1/ Major Machine Learning Applications

- Dynamic Pricing: traditionally, retail prices have been set based on some manual dynamic rules that consume a lot of energy and time for pricing and sometimes lead to wrong decisions. ML takes this process to another level by being able to process very large amounts of data and take into account various factors to make effective predictions as market conditions change. ML is considered powerful in dynamic pricing, as it can maximize income and achieve business goals.
- Speech Recognition: it is now embedded in everyday life, like Siri, Alexa, Google Assistant, Cortana, etc. Speech recognition is used in the workplace (scheduling meetings, printing documents, taking minutes), banking (payments, transactions, checking account balances), marketing, healthcare, language learning and much more. Speech recognition is also found on phones, smartwatches, tablets, and even it automates our homes.
- Agriculture: ML is used in agriculture to detect disease, predict crop quality, gain knowledge, and collect data to predict production of livestock. It is useful in this field because it can reduce the risk of farm or crop destruction by wild animals in a remote location by using programmed monitoring systems. It can also reduce labor costs by replacing human workers with AI harvesting robots and implementing an automated irrigation system that checks moisture, soil composition, and temperature. On the other hand, ML enables species recognition and breeding.

- Healthcare: ML is evolving with significant impact on the medical field, improving the efficiency of patient care in areas such as robotic surgery, drug discovery, smart health records, care for special-needs groups, personalized medicine, and many others. This helps in decision support systems to detect diseases and treatments and automate many tasks.

- Autonomous vehicles: a connected car drives without human assistance by identifying objects, analyzing situations, and making decisions. This is very helpful for people with disabilities and improves road safety by reducing accidents.

- Customer Support Chatbots: it uses natural language processing and ML, to simulate human conversations and respond to customer inquiries while reducing customer support costs. This process makes response time faster and better than the traditional customer agent routing process. It enhances the customer experience by being available 24/7.

- Traffic Prediction: used by most drivers to determine road closures, shortest route, and duration to destination. It predicts traffic volumes by checking congestion and vehicle movements to determine the optimal route. In some situations, it also checks the weather, as it can affect the road situation and driving speed.

- Real Estate Price Prediction: called “intelligent investing” to determine the best place and date to sell the house, set an optimal rental/sale price, or consider where/what home to buy to get the best financial return.

- Social Media Content Moderation: people use social media every day to share opinions and insights, respond to posts, express their feelings, communicate with others, etc. The freedom to post anything the user wants, anywhere, makes controlling offensive content a difficult task. Therefore, ML can detect obscene material such as violence or sexism, whether it is text, images, or videos that can be harmful to people of different ages or societies.

- Intelligent Video Surveillance: Used for person/object detection and face recognition. It can detect any abnormal event, such as a stranger entering a property, dangerous activities in public, and can send a real alert to emergency department while detecting criminal acts.

- Customer segmentation: is a marketing tool to understand the specific needs and interests of the target audience. Segmentation can be based on demographics, geographic location, behaviour, psychographics, or customer journey segmentation. This technique improves customer service by providing a better understanding of customer needs, focusing on the most profitable customers interested in that product, and leads to price optimization.

- Content/Product Recommendations: it displays the similar products/content that users have browsed, searched, rated or viewed by recording historical searches on a collaborative filtering system.

2.2/ APPLICATION OF ML FOR EMERGENCY RESPONSES

Various artificial intelligence techniques have been used in medicine and healthcare and have attracted attention in recent years. The emergency department is considered one of the most critical and important parts of any hospital, along with police, fire, and ambulance services. In the face of global population growth, traditional techniques for rapid and effective ED systems may not be sufficient, and the use of ML such as data mining, speech processing, classification, and clustering must be employed to improve the effectiveness of emergency responses. ML will obviously reduce human and material resources, leading to a reduction in the cost and time of providing ED services.

Different researches has promoted the use of ML in ER in hospitals: A notable study aims to make a transition from traditional emergency records to an electronic nursing report to help the emergency department with both data analysis and clinical use by providing important information that can change the management of this department. This method improved clinical care and quality assurance by integrating databases and registries [Redfield et al., 2020]. On the other hand, research attempted to place a fleet of ambulances at bases to increase the utility of the medical system's service level. An embedded simulator was integrated within a greedy allocation algorithm for a large Asian city and a significant result was demonstrated [Yue et al., 2012]. In addition, a study developed in [Kang et al., 2020] validated a deep-learning artificial intelligence algorithm to predict critical care needs during emergency medical services of a Korean national emergency department and outperformed prevailing triage tools and primitive warning scores. This was done by collecting information from 151 different ED in real time and authenticating run sheets from two different hospitals. On the other hand, a study collected 120600 administrative records from two hospitals in Ireland to predict the risk of patient admission to the ED allowing future resource planning and mitigation of an overloaded patient flow [Graham et al., 2018]. The results identified several criteria related to hospital admissions: hospital location, arrival mode, age, care group, triage category, and prior admission in the last month/year.

Similarly, ML has played a great role in ambulance, fire, and police emergency services. An AI decision support system has been developed [Sujan et al., 2022] to enable out-of-hospital cardiac arrest detection by ambulance service call centre operators so that immediate cardiopulmonary resuscitation can be initiated. Another re-

search [Okamoto Jr et al., 2020] proposed an intelligent traffic system that determines vehicle flow with minimal disruption to the traffic of other vehicles to minimize the arrival time of emergency vehicles such as police cars, fire trucks, and ambulances within the prescribed time to avoid fatalities or other complications.

To boot, many analyses and studies in the field of firefighters are noteworthy, where machine learning is used for various purposes. [Raveendran, 2020] looks at the future of intelligent firefighters using AI to reduce the risk of dangerous fire accidents and improve safety measures. Videos using virtual and augmented reality technology will help firefighters navigate panic situations.
Into the bargain, Fernandes, P.A.M predicted the fire spread in a flat terrain in Shrubland in Portugal on a series of experiments and prescribed fires in four different shrub fuel types considering weather, fuel conditions and fire spread rates up to 20 minutes [Fernandes, 2001]. Also, Pirklbauer, K. and Findling, R.D. proposed an approach for predicting the fire departments' deployment category based on time, weather, and location information using multiple machine learning algorithms [Pirklbauer et al., 2019]. On the other hand, Lian, X. et al. applied distributed computing and machine learning algorithms (Linear Regression, Decision Tree Regression, and Random Forest Regression) to predict the emergency response time for the San Francisco Fire Department [Lian et al., 2019].

Bradstock, R.A. et al. explored large fire ignition days probability in Sydney, Australia, using a Bayesian logistic regression influenced by the ambient and drought weather components of the Forest Fire Danger Index [Bradstock et al., 2009]. In addition, Coffield, S.R. et al. used decision trees to classify the final size of fire at the time of ignition in Alaskan boreal forests into small, medium, and large [Coffield et al., 2019]. Moreover, Fang, H. et al. implemented a machine learning based approach to identify automatically the stages of fire development in residential fires from a collection of fireground information using Gaussian Mixture Models and Hidden Markov Models [Fang et al., 2021]. Furthermore, O'Connor et al. developed a boosted logistic regression model to classify final fire locations using a dataset that includes topographic features, fuel types, and natural barriers to fire spread in southern Idaho and northern Nevada [O'Connor et al., 2017].

To come to the point, countless works have been published to optimize or analyze the problems of health care and emergency department services using Artificial Intelligence and Machine Learning.

2.3/ ML CHALLENGES

It is evident that ML and data mining have improved healthcare and emergency response in different ways, but many challenges in this area still exist:

1. Security and Privacy: security of healthcare data seems to be very critical for patients and organizations, as more and more security-related issues have emerged in recent years [Sharma et al., 2013]. During the ML process, large amounts of patient data can be shared, which can pose a threat to this confidential information. Some patients may not have the intention of disclosing their private health data for data mining extrapolations [Tomar et al., 2013]. In addition, it is not that patients are unwilling to disclose their data; it may be that the organization does not want them to. In general, it is a difficult process to collect data from different organizations [Ohno-Machado et al., 2015].

   Currently, various solutions are applied to data mining techniques to ensure privacy and security. Privacy Protected Data Mining (PPDM) [Vaidya et al., 2004, Nayak et al., 2011] such as data perturbation [Muralidhar et al., 2003, Kargupta et al., 2003], data suppression [Oliveira et al., 2004], and cryptography [Verykios et al., 2004] are among the techniques that restrict the data with certain rules and preserve confidentiality.

2. Data reliability: an important criterion in data mining is the quality of the data used in building the model to make predictions and for data analysis: good data leads to good decisions [Sharma et al., 2013]. The vast amounts of data collected by different providers can be incomplete, incoherent, noisy, and heterogeneous, mainly because healthcare organisations lack consensus and knowledge in using the metadata for data mining projects [Denaxas et al., 2016]. These problems could be caused by human errors or inaccuracies in the measurement of the data as well.

   Several actions can be taken to improve data quality to achieve better analysis, starting with the way data is collected, stored, and managed. It is important to regularly cleanse data by keeping useful, complete, and up-to-date information [Brownlee, 2020]. In addition, normalization, integration, and segmentation of data can also be used [Ali et al., 2014].

3. Data complexity: true dataset collected by real organisations are really heterogeneous. Such media data may include images, videos, natural language texts, spatial data, time series, audio data, and so on. Moreover, advances in medicine rely primarily on imaging technology, which increases the complexity of data [Hosseinkhah et al., 2009]. Conversely, there are many techniques to deal with such a problem, e.g., feature extraction [Guyon et al., 2008] and feature selec-

tion [Li et al., 2017].

4. Data modelling and interoperability: another important challenge in data mining is data modeling, which is a series of conceptual, logical, and physical processes to set up data and system behavior to meet business and healthcare needs [Blobel et al., 2018]. Furthermore, verifying the results of data mining is challenging because this decision-making can be subject to many fluctuations. Therefore, the results generated depend on the sensitivity and specificity of the data mining tool used, both of which affect the prediction result. In addition, the data mining results are not concrete and the interpretation of the data by non-health experts may lead to errors in decision making [Werts et al., 2000]. Hence, several tactics can be used for better data modeling: using structured data, verified models, training and testing a sufficient number of data points, and using a biased model during the machine learning process.

2.4/ CONCLUSION

In this chapter, an overview of various applications of machine learning in the field of emergency response, whether in hospitals, ambulances, police, or firefighters, was given. Data issues are of great importance to ensure good analysis and prediction results, therefore, the challenges of ML were addressed in the second part with different solutions.

3

TIME SERIES FORECASTING

3.1/ GENERAL OVERVIEW

This chapter illustrates the features of time series forecasting that qualify this work. Then, the related work of the fire dataset used in all experiments is presented, and finally, all the ML algorithms that were used in the different contributions and the statistical metrics are discussed.

3.1.1/ CHARACTERISTICS OF TIME SERIES

Time series forecasting (TSF), an important part of ML, consists of a sequence of data points set at a specific time interval, such as 1 hour, 1 day, 1 month, etc. In TSF, trends and seasonal variations are traced to examine and analyse historical observations to make better predictions for the future. The main features and characteristics of TSF are represented in the following part.

- Trends: refer to the increasing or decreasing motion of a time series. Depending on the time period over which the trend is observed, it may be relatively higher or lower values in the long or short term. There are two main types of trends: cyclical and seasonal. Cyclical trends repeat, but not necessarily in a given period of time, e.g., business cycles. In a cyclical trend, data does not fall and rise in a given period of time. Reciprocally, human activity can be expected to increase each year during the holidays. This is what defines a seasonal trend, also known as periodic. In other words, the seasonal trend involves rhythmic forces that operate in a periodic and regular manner.

- Random fluctuations: it concerns irregular events that occur during a time series observation. Such random variations are usually uncontrollable and unpredictable, but can be determined by applying mathematical procedures.

- Stationarity: TSF can also be stationary if the static properties do not change over time, meaning that the data values are independent of time. This can be seen as a constant metric such as variance and mean. In contrast, non-stationary time series are random processes that do not have a constant distribution over time.

To put it all in a nutshell, TSF is considered a reliable technique because it collects information over a period of time, such as a day, month, year, etc. On the other hand, the seasonal pattern is an important feature of the TSF that helps to understand, analyze, and predict future observations that depend on historical patterns. In short, the TSF is an excellent tool for assessing rising and falling trends as well as measuring the growth of an organisation, helping to make structured decisions about the future.

3.1.2/ TIME SERIES FORECASTING PROGRESS

In recent years, many influential works on time series forecasting have been published, with enormous progress in this field. Gardner and Snyder pioneered forecasting when they published two papers on exponential smoothing methods in the same year (1985) in the field of time series forecasting. Gardner provided a review of all the existing work done to that date and extended his research to include a damped trend [Gardner Jr, 1985].

After Gardner's work, Snyder demonstrated that simple exponential smoothing could arise from state space model innovation [Snyder, 1985]. Most of the researches since 1985 has focused on empirical properties [Bartolomei et al., 1989], forecasts evaluations [Sweet et al., 1988], and proposition of new methods for initialization and estimation [Ledolter et al., 1984]. Many studies have stimulated the use of exponential smoothing methods in various areas, such as air passengers [Grubb et al., 2001], computer components [Gardner Jr, 1993], and production planning [Miller et al., 1993]. Later on, numerous variations on original methods have been proposed to deal with continuities, constraints, and renormalization at each period time [Williams et al., 1999], [Rosas et al., 1994]. Multivariate simple exponential smoothing was used for processing the control charts by introducing a moving average technique [Gang et al., 2013]. Moreover, [Taylor, 2003], and [Hyndman et al., 2002] have extended basic methods and have included all 15 different exponential methods. They have proposed models that correspond to multiplicative error cases and additive errors. However, these methods were not unique since it has been known that ARIMA models give equivalent results in forecasting, but the innovation in their work was that statistical models can lead to non-linear exponential smoothing methods.

Early studies of time series forecasting in the nineteenth century were globally based on the idea that every single time series can be seen as a realization of a stochastic process. Based on this simple proposal, many time series methods since then have been

developed. Researchers such as Walker, Yaglom, Slustsky, and Yule [Chen et al., 2014] formulated the concept of moving average MA model and autoregressive AR models. The concept of linear forecasting came out by the decomposition theorem. After that, many studies have appeared dealing with parameter identification, forecasting estimation, and model checking.

3.2/ APPLICATION OF TIME SERIES TO THE FIREMEN DATASET

This section summarizes a literature review of previous studies on the same topic of firefighters' interventions using the same dataset as this dissertation.

The study began in mid-2019 when [Ñahuis et al., 2019] investigated that firefighters' missions are predictable, to which artificial intelligence technology can be applied. They used long short-term memory neural networks to predict 2017 deployments from those of 2012-2016. Then, [Couchot et al., 2019] used extreme gradient boosting on anonymized data to predict the number and type of firefighters' deployments by civil protection services. Also, in [Guyeux et al., 2019], explanatory variables were added based on calendar, weather, road traffic, astronomical data, etc., and the learning process was performed on an ad hoc multilayer perceptron to predict firemen operations for 2017. In [Arcolezi et al., 2020], on the other hand, the researchers noted that it was essential to anonymize the data using Differential Privacy to avoid information leakage with such sensitive data. XGBoost was implemented to generate the forecast.

In [Cerna et al., 2020], predictions for firemen interventions were compared using XGBoost and LSTM, showing that machine learning can produce feasible predictions even for rare events such as natural diseases. In addition, [Cerna et al., 2020] has developed indicators for detecting breakdowns caused by the temporal state of human and vehicle materials to increase operational resilience and improve the efficiency of fire department responses.
Research was also conducted on a study of fireman interventions during the COVID-19 period: In [Cerna et al., 2021], the impact on ambulance turnaround time was analyzed. The number of service failures was calculated, resulting in a decision support tool by determining ambulance dispatch times for medical and personal services.

Finally, [Guyeux et al., 2022] proposed an operational knowledge base with relevant and updated content aimed at industrializing this process, leading to an improvement in the operational response of emergency services.

3.3/ Machine Learning algorithms applied

A variety of ML and time series algorithms were applied in this dissertation and are presented in the following sections.

3.3.1/ Auto Regression

Auto regression (AR) is a statistical time series model that predicts an output for the near-future (number of houses sold, price of something, number of intervention,...) based on past values. It was originated in 1920 by Udny Yule, Eugen Slutsky, and Ragnar Frisch [Klein et al., 1997]. For instance, to predict today's value based on yesterday, last week, last month, or last year's data points. AR models are also called Markov models, conditional models, or transition ones.

Regression uses external factors which are independent as an explanatory variable for the dependent values. Autoregression model is conditioned by the product of certain lagged variables, and coefficients allowing inference to be made. In reality, AR works poorly if the future predictors are unknown because it requires a set of predictor variables. On the other hand, AR is capable of adjusting the regression coefficient β and violating the assumption of uncorrelated error since the independent observations are time lagged values for the dependent observations.

In an AR model, the value of the predicted outcome variable (Y) at some time t is:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t \tag{3.1}$$

where the parameters $\beta_0 + \beta_1 Y_{t-1}$ rely on the past and ϵ_t which is the white noise could not be predictable from the past. It is important to mention that knowing the previous lagged values of Y_{t-2}, Y_{t-3} does not affect the prediction of Y_t because as shown in the formula, Y_t is affected only by Y_{t-1}.

3.3.2/ Moving Average

Moving Average MA is a model introduced in 1921 by Hooker that considers multiple period averages to predict future output and event [Hooker, 1921]. It is an effective and naive technique in time series forecasting, used for data prediction, data preparation, or feature engineering. It uses the most recent historical data values to generate a forecast. MA removed the fine-grained variation between time steps to expose the signal.

This method uses the average data period's number. The term "moving" indicates the up and down moves of the time series made to calculate the average of a fixed number

of observations. On the other hand, the process of averaging relies on the overlapping observations that create averages. Moving Average method can be used for both linear and non-linear trends. However, it is not applicable for short-time series forecasting fluctuations because the trend obtained by applying the model is neither a standard curve nor a straight line. Besides, trend values are not available for some intervals at the start and end values of a time series.

The outcome value in the MA(q) model, a moving average model of order q, is presented as the following:

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q} \tag{3.2}$$

where ϵ_t is the white noise. This technique involves creating a new time series with compromised values of row observations, and average in the original dataset time series. Also, it relies on past forecast errors.

3.3.3/ AutoRegressive Integrated Moving Average

ARIMA, also called Box-Jenkins, is a model proposed by George Box and Gwilym Jenkins in 1970 by using a mathematical approach to describe changes in the time series forecasting [George, 1970]. ARIMA is an integration of autoregression and moving average methods that use a dependent relationship between an observation and some number of lagged observations by differencing between raw observations. It subtracts an observation from the previous time step and takes into consideration the residual error. ARIMA is a powerful model as it takes into consideration history as an explanatory variable, but in such a model, the data cost is usually high due to the large number of observations needed to build it properly. A standard notation for ARIMA being used is ARIMA (p,d,q) where:

- p: is the auto-regressive part of the model, which means the number of lag observations that are included into the model. It helps to incorporate the effect of past values of the model. In other words, it is logical to state that it is likely to need 5 firefighters tomorrow if the number of interventions was 5 for the past 4 days. A stationary series with autocorrelation can be corrected by adding enough AutoRegression terms.

- d: is the integrated part of the model. It shows the degree of difference by the number of times that the raw observations have been different. This is similar to stating that if in the last 4 days the difference in the number of interventions has been very small, it is likely to be the same tomorrow. The order of differentiation required is the minimum order needed to get a near-stationary series.

- q: order of moving average, which is the size of the MA window. The Autocorrelation

graph shows the error of the lagged forecast. The ACF shows the number of MA terms required to remove autocorrelation in the stationary series.

3.3.4/ PROPHET

A prophet forecasting model is an open-source algorithm designed by Facebook in 2017 [Taylor et al., 2018] for time series having common features and intuitive parameters, where experts and non-experts in statistics and time series forecasting can use it. The Prophet is based on time series models and is dependent on four major components: (1) yearly and weekly seasonality; (2) non-linear trend; (3) holidays; and (4) error. The Prophet fits very well for data that has at least one year of historical inputs with daily periodicity. It is very fast in terms of fitting the model, working without converting data into time-series objects, and being robust to missing values. In addition, Prophet is simpler compared to other time series forecasting algorithms because it requires fewer parameters and models. The Prophet works as follows:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{3.3}$$

where:

- g(t): describes the increase or decrease trends in the long-term data.
- s(t): represents the impact of seasonal factors over the year on the time-series data.
- h(t): models how large events and holidays affect the data.
- ϵ_t: shows the non reducible error term.

3.3.5/ EXPONENTIAL SMOOTHING

There are mainly three types of Exponential Smoothing: single (does not treat systematic structure), double (treats trend), and triple (both trend and seasonality).

3.3.5.1/ SIMPLE EXPONENTIAL SMOOTHING (SES)

SES was developed by Robert G. Brown (1956) [Brown, 1956] and is naturally called Single Exponential Smoothing. It is a time series forecasting method for data with no clear trend or seasonal pattern. Forecasts are calculated by using the weighted average of the previous level and the current observations. SES associates more weightage to recent observations and fewer weights to older ones [Gardner Jr, 1985] and essentially requires

a “level” component called alpha (α). The formula for Simple Exponential Smoothing is as follows:

$$S_t = \alpha y_t + (1 - \alpha) S_{t-1}, t > 0 \tag{3.4}$$

where:

- α: smoothing coefficient between 0 and 1
- S_t: forecast value for period t
- y_t: refinement constant for the whole data

3.3.5.2/ Double Exponential Smoothing (Holt)

Charles C. Holt (1957) extended Single Exponential Smoothing to allow prediction of data with trend [Holt, 2004]. Holt’s method assumes that datasets have a trend and do not have seasonality, so it uses two components, “level” and “trend”. This method is mainly used for linear trends with short to medium forecast periods. The two smoothing parameters for level and trend, alpha (α) and beta (β) respectively, are ranged between 0 and 1. A second equation is added to handle the trend aspect:

$$b_t = \beta(S_t - S_{t-1}) + (1 - \beta) b_{t-1} \tag{3.5}$$

wherein:

- β is the refinement constant for trends
- b_t is the trend for period t

3.3.5.3/ Triple Exponential Smoothing (Holt-Winters)

Holt-Winters (1960) is an extension of Holt’s method by Holt’s student, Peter Winters, who assumes the existence of seasonal and trend variations for data observation [Winters, 1960]. It considers level, trend, and seasonality with the corresponding parameters alpha (α), beta (β), and gamma (γ). Triple Exponential Smoothing adds a third equation to the single and double smoothing as follows:

$$I_t = \gamma y_t + (1 - \gamma) I_{t-l} + m \tag{3.6}$$

where:

- γ is the seasonal smoothing constant

- I_t is the seasonal index for period t
- l is the size of the season

3.3.6/ eXTREME GRADIENT BOOSTING

XGBoost was developed in 2016 by Tianqi Chen [Chen et al., 2016], and is one of the most powerful and leading machine learning algorithms in the field of ensemble learning, implementing gradient boosted decision trees. XGBoost is faster and more accurate compared to other gradient boosting methods as it builds a strong sequence model from weak models, following the concept of level-wise growth shown in Figure 3.1.

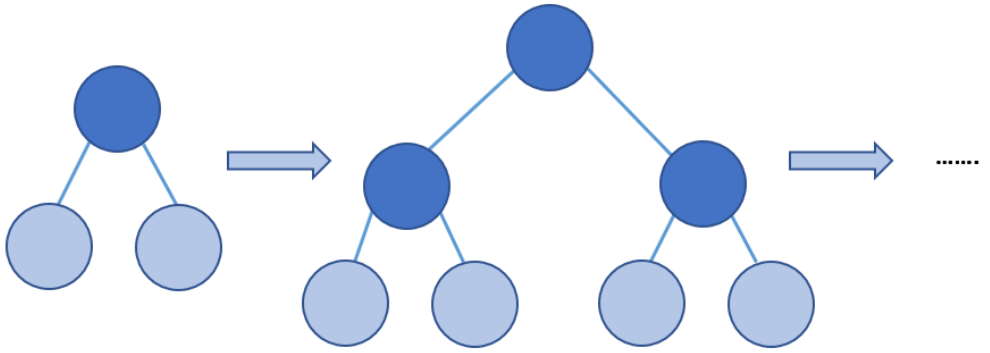


Figure 3.1: Level tree growth strategy

It is the improved Gradient Boosting Decision Tree algorithm [Friedman, 2001] that considers an objective function to prevent over-fitting and can automatically use the CPU for parallel multi-threaded computations to improve the speed and performance of the model. It considers the leaves of the current decision tree and checks for the possibility of improving the model when the leaf is transformed into a new "if" statement. The more "if" conditions are added to the model, the stronger the model becomes. XGBoost is represented in the following formula:

$$\hat{y}_i = \sum_{t=1}^{n} f_t(x_i) f_t \in F \tag{3.7}$$

where:

- n: presents the number of trees
- $\hat{y}_i$: is the predicted value
- x_i: represents the ith sample of the input
- n: is the number of trees

- F: is the set of all possible CART
- f_t: expressed the function in the function space

3.3.7/ LIGHT GRADIENT BOOSTING MACHINE

LightGBM, an extension of XGBoost was developed by Microsoft specifically by Guolin Ke [Ke et al., 2017] in 2016 and is similar to XGBoost, which uses decision trees for classification and regression. The term “Light”, after which this algorithm is named, refers to the computational power that leads to faster results and the ability to deal with metadata that consumes less memory. Unlike XgBoost, LightGBM solves predictions based on a tree’s leaf-wise growth strategy, which can lead to an over-fitted model as illustrated in Figure 3.2.

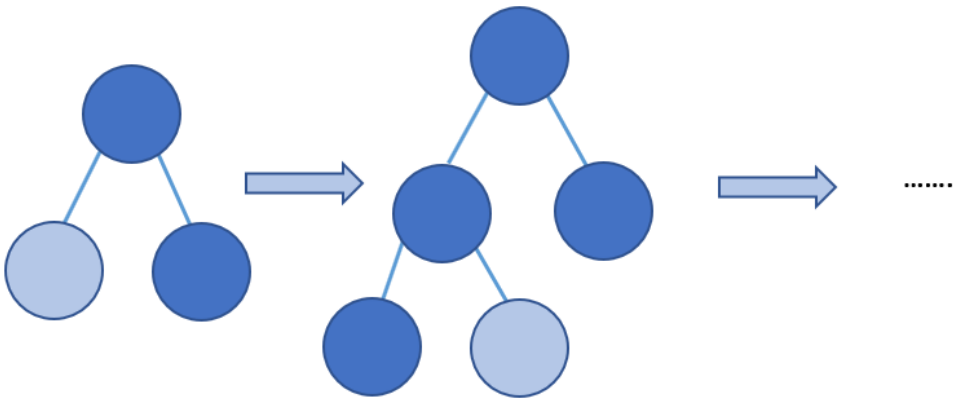


Figure 3.2: Leaf tree growth strategy

3.3.8/ LONG SHORT-TERM MEMORY

The LSTM, which belongs to Deep Learning and is a kind of recurrent neural network (RNN), was developed by Juergen Schmidhuber in 1997 [Hochreiter et al., 1997]. The LSTM memory block contains activation functions, weight inputs, inputs from various recurrently connected blocks called memory blocks, and outputs. The name LSTM indicates that the program uses short-term memory to create longer memories. This algorithm was designed to solve the problem of RNN in terms of long-term dependence. LSTM has three gates that play the roles of filters: Forget, Input, and Output.

Each gate is exhibited as the following:

1. Input:

$$i_t = \delta(w_i[h_{t-1}, x_t] + b_i) \tag{3.8}$$

where:

- i_t: input gate
- δ: sigmoid function
- w_i: weight for the input neuron
- h_{t-1}: output of the LSTM block of the previous time t-1
- x_t: input at the current time
- b_i: bias for the input gate

2. Forget:

$$f_t = \delta(w_f[h_{t-1}, x_t] + b_f) \tag{3.9}$$

where:

- f_t: forget gate
- w_f: weight for the forget neuron
- b_f: bias for the forget gate

3. Output:

$$o_t = \delta(w_o[h_{t-1}, x_t] + b_o) \tag{3.10}$$

where:

- o_t: output gate
- w_o: weight for the output neuron
- b_o: bias for the output gate

3.4/ Statistical metrics and frameworks

Various methods were used in the experiments, whether to calculate specific metrics or to test predictive accuracy using different ML algorithms.

3.4.1/ Mean Absolute Error

MAE measures the average difference in the absolute values of the prediction errors in a set of forecasts for all instances. Each prediction error marks the difference between the actual value and the calculated value. MAE is represented in the formula:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - x_j| \tag{3.11}$$

where:

- n: number of all the values
- y_j: actual value for the ith observation
- x_i: predicted value for the ith observation

3.4.2/ Root mean squared error

RMSE measures the average magnitude of the prediction errors, which indicates the absolute fit of the model to the dataset. RMSE is depicted in the formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - x_j)^2} \tag{3.12}$$

3.4.3/ Silhouette score

It is used to evaluate the quality of clusters created after applying clustering algorithms by calculating the gaps between each data sample of the same cluster. It ranges from -1 to 1, and the closer the value is to 1, the better the clusters are separated from each other and the denser they are. A value of 0 reflects overlapping clusters, and a negative value means that the samples may have been incorrectly assigned to the wrong clusters. The silhouette coefficient is calculated using the following formula:

$$S = \frac{(p-q)}{max(p,q)} \tag{3.13}$$

where:

- p: mean nearest-cluster distance, which is the mean distance between the observations and all points of the nearest cluster
- q: mean intra-cluster distance to all the points in the same cluster

3.4.4/ Optuna

It is an automatic hyperparameter tuning framework developed for ML and used in almost all experiments to tune the hyperparameters to optimize the accuracy of the applied algorithms. Optuna uses historical experimental protocols to verify the coming values of the hyperparameters under test, regulating their configurations.

3.4.5/ Feature Selection

Using non-essential features in a dataset to train a model can cause the model to learn from noise and slow down the process. Including all columns in the training process does not significantly affect the prediction and accuracy of the classifier. Consequently, feature selection is the method to get rid of irrelevant attributes, called noise and keep only the useful input. Moreover, this technique helps to reduce the size of the dataset, which leads to a reduction in training time, computational resources, and complexity of the overall model.

3.5/ Conclusion

Since this dissertation deals specifically with the topic of time series forecasting, an overview of its characteristics has been provided in this chapter, along with a state of the art of previous studies by researchers who have worked with the exact same dataset. In addition, this chapter discusses all of the machine learning techniques used in this dissertation and explains the statistical metrics used to calculate the hyperparameters and optimize the accuracy of the techniques.

4

TECHNICAL TOOLS

In order to perform the various experiments of this dissertation and implement the several machine learning algorithms, numerous technical tools were used to achieve the goal. Therefore, this chapter includes all the infrastructure utilized, such as hardware, software, packages, libraries, and frameworks.

At the beginning of the work, the experiments were performed on a 2.7 GHz Core i7 processor with 8 GB RAM, then the hardware was replaced by an 11th Gen Intel Core i7, 3.00GHz with 16 GB RAM. It is obvious that with the increase of RAM, the processing speed has incremented, as the speed of information transfer has augmented.

4.1/ PLATFORM AND LANGUAGE

Jupyter is the interface used in this work. It is an interactive and flexible platform for configuring and writing codes for data science and machine learning. Moreover, it is very easy to share a notebook created in the Jupyter environment, as it is a free open source web tool.

The programming language used in all experiments is Python, which is supported by the Jupyter notebook. Python is an object-oriented language that includes a large number of machine learning frameworks and libraries that certainly make the code simple and reduce the time needed for complex implementations. Python is also a very simple language and fast compared to many other programming languages. In addition, Python supports data visualisation, which is crucially beneficial tool for data scientists.

4.2/ PACKAGES AND LIBRARIES

Libraries and packages are essential in ML to perform various tasks such as data cleansing, normalization, visualization, inspection, save/load, statistical analysis and much

more. All of the following packages, libraries, and modules have been imported into the Jupyter notebook to perform specific tasks.

- NumPy: stands for Numerical Python and is the basic package used in data science to perform numerical calculations. It provides an array object that is much faster than a normal Python list. In this work, the whole infrastructure was installed with the pip command in Jupyter, example: *pip install numpy*
- Pandas: built on top of NumPy and is mainly used for data manipulation and analysis. Pandas works with datasets and helps draw conclusions based on statistical theories through data cleaning and manipulation.
- Matplotlib: used for data visualization, animation, graphic plots, and diagrams.
- Scikit-learn (Sklearn): a powerful library designed specifically for ML. It can handle supervised, unsupervised learning, clustering, data reduction, cross-validation, feature selection, and other tools via an interface in Python.
- Pathlib: works with paths and directories, e.g. when reading or loading a file.
- Tslearn: toolkit used especially for time series analysis.
- Statsmodels: a module that contains functions and classes for creating statistical models and running tests.
- Seaborn: used for semantic mapping and the creation of informative graphs and charts.
- Ruptures: detects breakpoints and provides methods for data segmentation.
- Pycaret: automates ML workflows such as feature engineering, tuning hyperparameters, replacing missing data, and more.
- Math: used for mathematical functions.
- Fbprophet: used to implement the prophet open-source library

On top of that, many other libraries based on the imported packages have been used to make predictions, deal with statistics, or conduct data analysis, either theoretically or visually. Amongst: timeSeriesKMeans, train_test_split, mean_squared, mean_absolute_error, LinearRegression, XGBoost, LSTM, LightGBM, feather, MinMaxScaler, SimpleExpSmoothing, Holt, ExponentialSmoothing, EarlyStopping, dateFormatter, autocorrelation_plot, timeSeriesSplit, plot_acf, seasonal_decompose, prophet, datetime, warnings and many more as shown in Table 4.1.

Packages			
Cython	bz2	mmapfile	statsmodels
IPython	cProfile	mmsystem	string
PIL	calendar	modulefinder	stringprep
StandardScaler	cffi	msgpack	struct
__future__	cgi	msilib	subprocess
_abc	cgitb	msvcrt	sunau
_aix_support	chunk	multiprocessing	symtable
_argon2_cffi_bindings	click	nbclient	sys
_ast	cloudpickle	nbconvert	sysconfig
_asyncio	cmath	nbformat	tabnanny
_bisect	cmd	nest_asyncio	tarfile
_blake2	code	netbios	tblib
_bootsubprocess	codecs	netrc	telnetlib
_bz2	codeop	nntplib	tempfile
_cffi_backend	collections	notebook	terminado
_codecs	colorama	notebooks	test
_codecs_cn	colorsys	nt	testpath
_codecs_hk	commctrl	ntpath	textwrap
_codecs_iso2022	compileall	ntsecuritycon	this
_codecs_jp	concurrent	nturl2path	threading
_codecs_kr	configparser	numba	threadpoolctl
_codecs_tw	contextlib	numbers	time
_collections	contextvars	numpy	timeit
_collections_abc	copy	odbc	timer
_compat_pickle	copyreg	opcode	tkinter
_compression	crypt	operator	tlz
_contextvars	csv	optparse	token
_csv	ctypes	os	tokenize
_ctypes	curses	packaging	toolz
_ctypes_test	cycler	pandas	tornado
_datetime	cython	pandocfilters	trace
_decimal	dask	parso	traceback
_distutils_hack	dataclasses	partd	tracemalloc
_elementtree	datetime	pathlib	traitlets
_functools	dateutil	patsy	tslearn
_hashlib	dbi	pdb	tty
_heapq	dbm	perfmon	turtle
_imp	dde	pickle	turtledemo
_io	debugpy	pickleshare	types

_json	decimal	pickletools	typing
_locale	decorator	pip	unicodedata
_lsprof	defusedxml	pipes	unittest
_lzma	difflib	pkg_resources	urllib
_markupbase	dis	pkgutil	uu
_md5	distributed	platform	uuid
_msi	distutils	plistlib	venv
_multibytecodec	doctest	poplib	warnings
_multiprocessing	email	posixpath	wave
_opcode	encodings	pprint	wcwidth
_operator	ensurepip	profile	weakref
_osx_support	entrypoints	prometheus_client	webbrowser
_overlapped	enum	prompt_toolkit	webencodings
_pickle	errno	pstats	wheel
_py_abc	executing	psutil	widgetsnbextension
_pydecimal	faulthandler	pty	win2kras
_pyio	filecmp	pure_eval	win32api
_pyrsistent_version	fileinput	pvectorc	win32clipboard
_queue	fnmatch	py_compile	win32com
_random	fontTools	pyclbr	win32con
_sha1	fractions	pycparser	win32console
_sha256	fsspec	pydoc	win32cred
_sha3	ftplib	pydoc_data	win32crypt
_sha512	functools	pyexpat	win32cryptcon
_signal	gc	pygments	win32event
_sitebuiltins	genericpath	pylab	win32evtlog
_socket	getopt	pyparsing	win32evtlogutil
_sqlite3	getpass	pyrsistent	win32file
_sre	gettext	pythoncom	win32gui
_ssl	glob	pytz	win32gui_struct
_stat	graphlib	pywin	win32help
_statistics	gzip	pywin32_bootstrap	win32inet
_string	hashlib	pywin32_testutil	win32inetcon
_strptime	heapdict	pywintypes	win32job
_struct	heapq	pyximport	win32lz
_symtable	hmac	qtconsole	win32net
_testbuffer	html	qtpy	win32netcon
_testcapi	http	queue	win32pdh
_testconsole	idlelib	quopri	win32pdhquery

_testimportmultiple	imaplib	random	win32pdhutil
_testinternalcapi	imghdr	rasutil	win32pipe
_testmultiphase	imp	re	win32print
_thread	importlib	regcheck	win32process
_threading_local	inspect	regutil	win32profile
_tkinter	io	reprlib	win32ras
_tracemalloc	ipaddress	rlcompleter	win32rcparser
_uuid	ipykernel	runpy	win32security
_warnings	ipykernel_launcher	ruptures	win32service
_weakref	ipython_genutils	sched	win32serviceutil
_weakrefset	ipywidgets	scipy	win32timezone
_win32sysloader	isapi	seaborn	win32trace
_winapi	itertools	secrets	win32traceutil
_winxptheme	jedi	select	win32transaction
_xxsubinterpreters	jinja2	selectors	win32ts
_yaml	joblib	send2trash	win32ui
_zoneinfo	json	servicemanager	win32uiole
abc	jsonschema	setuptools	win32verstamp
adodbapi	jupyter	shelve	win32wnet
afxres	jupyter_client	shlex	winerror
aifc	jupyter_console	shutil	winioctlcon
antigravity	jupyter_core	signal	winnt
argon2	jupyterlab_pygments	site	winperf
argparse	jupyterlab_widgets	six	winpty
array	keyword	skelm	winreg
ast	kiwisolver	sklearn	winsound
asttokens	lib2to3	smtpd	winxpgui
asynchat	linecache	smtplib	winxptheme
asyncio	llvmlite	sndhdr	wsgiref
asyncore	locale	socket	xdrlib
atexit	locket	socketserver	xml
attr	logging	sortedcontainers	xmlrpc
attrs	lzma	soupsieve	xxsubtype
audioop	mailbox	sqlite3	yaml
backcall	mailcap	sre_compile	yellowbrick
base64	markupsafe	sre_constants	zict
bdb	marshal	sre_parse	zipapp
binascii	math	ssl	zipfile
binhex	matplotlib	sspi	zipimport
bisect	matplotlib_inline	sspicon	zlib
bleach	mimetypes	stack_data	zmq
bs4	mistune	stat	zoneinfo
builtins	mmap	statistics	~umpy

Table 4.1: Modules, libraries and packages imported

4.3/ CONCLUSION

The popularity of using Python in machine learning and data science is vast due to the large collection of libraries, modules, and frameworks that can support it. Regardless of the type of data, Python libraries provide flexible functionality that can be used to accomplish any task with great efficiency. All libraries are open-source, platform-independent, free, and easy to learn. Moreover, their use reduces debugging time and coding complexity.

III

CONTRIBUTIONS

This part presents the main contributions of this work using Time Series Forecasting and various ML techniques. The datasets used are explained, related work on each topic is examined, experiments are performed, and the results are presented. A discussion is given for each contribution to analyze the results of the forecasts.

5

FORECASTING THE NUMBER OF FIREMEN INTERVENTIONS

Time series forecasting is one of the most attractive analyses of datasets that involve a time component to extract meaningful results in the economy, biology, meteorology, civil protection services, retail, etc. In this chapter various types of algorithms for predicting firemen operations were applied, such as Autoregression, Moving Average, Autoregressive Integrated Moving Average, Prophet, as well as three different time series exponential forecasting algorithms. Optimal values were selected, algorithms were implemented and then compared between applied models on the same dataset. The database used includes hourly recorded deployments in Doubs, France.

5.1/ INTRODUCTION

Many studies show that achieving good forecasts is vital in many activities. The fact of gathering a collection of observations over time will provide predictions of new observations in the future and extract meaningful characteristics of the data and statistics in different time intervals: hours, days, weeks, months, and years. Due to the French economic crisis (closure of small hospitals, population growth, etc.), the impact of optimizing the number of human interventions leads directly to a reduction and better control of financial, human, and material resources. This would also have a remarkable impact on protecting people, the environment, and property from damage, critical incidents, and disasters.

The usage of data science, machine learning, and time series forecasting is feasible in the prediction of firefighters' interventions since it is logical to assume that firefighters' interventions are affected somehow by temporal, climatic, and other events such as new year's eve, snowfall, traffic peak times, fires in summer, holidays, etc. Following

this principle, it will be possible to analyze past observations using historical values and associated patterns to predict future deployments. Therefore, these properties are well represented in time series forecasting approaches such as AR, MA and ARIMA as well as Exponential Smoothing, which consider the trend, level, and seasonality of a time-ordered series. They analyze and forecast data observations, and the result can lead to better decisions, as it is reasonable that the number of deployments in the coming hour tends to be influenced by the number of deployments in the previous hour.

The goal of this contribution is to size the number of firefighters according to the need and demand by predicting the firemen resources over short-term and long-term period; a greater number of firefighters should be available when they are most used. Indeed, the number of guards available should be related to the location, number, and type of the intervention. For example, during the weekend, when accidents actually increase, the number of firefighters ready to serve society should be greater than on a regular working weekday where most of the people reside in their offices. A trusted result and high prediction are affected by many factors; the algorithm used to train and test the dataset plays a big role, as well as the chosen parameters.

The dataset used in AR, MA, ARIMA , and Prophet contains information about firefighters' interventions in the Doubs-France region from "01/01/2006 00:00:00" to "31/12/2017 23:00:00". However, for the exponential smoothing methods, the dataset was changed to include only the period from "January 1, 2015 00:00:00" to "December 31, 2019 23:00:00". All the data was registered by the fire and rescue department SDIS 25 by blocks of one hour [Couchot et al., 2019]. An overview of the number of firefighters' interventions through the years is shown in Figure 5.1. Statistical parameters are calculated for different Machine Learning algorithms applied to predict the number of firefighters' interventions. The remainder of this chapter is structured as follows: the literature review provides an overview of related work done by researchers on the same topic; parameters chosen, showing the values used for the corresponding algorithms; building data with Prophet by applying this Facebook tool to the firefighters' dataset; obtaining results; interpreting results; and conclusion.

5.2/ Literature Review

Several studies conducted by researchers focus on the application of AR, MA and ARIMA for time series analysis for different purposes to predict the observation at time (t+1) considering historical data for the same observation. A number of studies related to these techniques are highlighted below.
In their work [Chujai et al., 2013], the authors build a model to predict household electric-

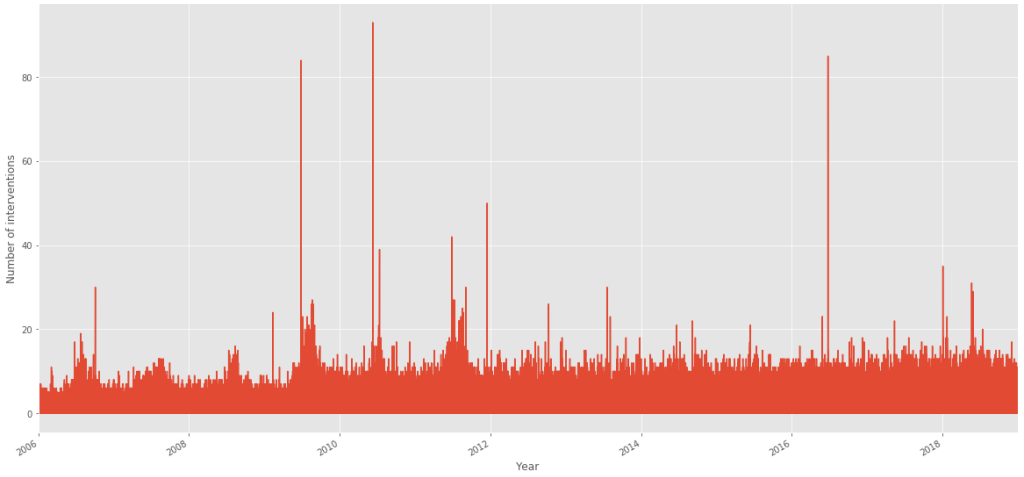


Figure 5.1: Number of firefighter interventions from 2006 until 2017.

ity consumption from December 2006 to November 2010 by applying ARIMA and finding the most appropriate time to forecast in daily, weekly, monthly or quarterly. Another work achieved by [Devi et al., 2013] aims at predicting stock trends in order to improve investment decisions. To this end, an overview of marketing investments was constructed using historical data over a five-year period for various companies. The model was trained using different algorithms including ARIMA with various parameters.
On the other hand, [Zhang et al., 2017] proposed a wavelet ARIMA/ARMA model to predict the short-time series of particulate matter concentrations with an aerodynamic diameter below 10 micrometers (PM10) from 4 stations in Taiyuan, China. Such prediction is difficult because of the uncertainty in the meteorological and emission description fields. Moreover, [Kumar et al., 2014] predicted sugarcane production in India for the five-year period using ARIMA. Parameters p, d, and q were selected to fit the model optimally using partial autocorrelation coefficients of residuals (PACF) and autocorrelation function (ACF). In addition, [Voyant et al., 2020] proposed a functional model based on a periodic autoregressive model to predict the solar irradiation, intending to improve the integration of intermittent solar systems.

[Özgür, 2005] compared numerical and graphical AR with neural network model predictions to predict 7- and 4-year flow in 2 rivers in the United States. This study can help with agricultural water management, avoid water shortages, and reduce the possibility of flood damage. Furthermore, in [Chen et al., 2008], a short-term property crime forecast using ARIMA for a city in China was performed. This work demonstrated an efficiency in decision-making for the government, emergency management, and police stations. The fitting and results were compared with SES and Holt exponential smoothing. On the other hand, a study presented in [Jadhav et al., 2017] predicts the prices of three crops in the Indian state of Karnataka using data from 2002 to 2016 using the ARIMA technique. Paddy, ragi, and maize are considered the most important crops as they are included

in the majority of food products and account for 40% of grain production in Karnataka. This prediction will provide farmers with a strategy to produce and market these crops by getting better prices and increasing their income.

In the bargain, the application of Exponential Smoothing techniques showed a great impact in data science over the years.

Singh, K. et al. [Singh et al., 2019] implemented in his research Exponential Smoothing method to predict the number of tourists for 2018 of an Indian state using the Java programming language for the years 2008 to 2017. He applied different values of smoothing constant to find the best accuracy of the model. Zafar S. et al. [Saba Zafar et al.,] analyzed and studied temperature data and variability of two major regions using the Simple Smoothing Technique and concluded that SES gives the best predictive values compared to other models.

Moreover, Argawu, A. [Argawu, 2020] predicted the number of COVID-19 new cases in the 10 most infected African countries by applying regression, ARIMA, and Exponential Smoothing Models. Yasar, H. and Kilimci, Z.H. [Yasar et al., 2020] emphasized how to mix Time Series Forecasting methods with Financial Sentiment Analysis data collected from Twitter, Instagram, and Facebook. In their case study, they employed ARIMA, Holt, and Holt-Winters to provide a more consistent exchange rate prediction to any user wishing to exchange Turkish Lira /US dollars, and they ended their study with the best-observed performance belonging to the Holt-Winters' method.

Anggrainingsih, R. et al. [Anggrainingsih et al., 2015] analyzed time series data of website visitors using the Triple Exponential Smoothing method. Their results showed the optimal alpha, beta, and gamma for the best prediction accuracy. Lai, K.K. et al. [Lai et al., 2006] proposed a hybrid methodology by integrating Neural Network with Exponential Smoothing for financial time series prediction. Their experimental results considered the accuracy and directional predictions and showed that the hybrid-integrated method performs better than the two benchmark models. Jones, S.S. et al. [Jones et al., 2008] examined and evaluated the use of SARIMA, Exponential Smoothing, time series regression, and ANN to predict daily patient volume in the emergency department, compared the results with the multiple linear regression model previously performed, and concluded that the regression-based model provided the most consistent accuracy.

5.3/ DATASET

The dataset used in this chapter contained information on firefighters from 2006 to 2018 collected by the SDIS 25 fire and rescue department in the Doubs region, France. Nev-

ertheless, the dataset was pruned for the exponential smoothing experiments to reduce computational effort and resources without affecting the prediction results. The data is sorted by date in ascending order and contains 119 columns, including the index, containing various geographic, temporal, and spatial information and, most importantly, the number of interventions of firemen for each specific date. The main attributes that provide information about the date are hour ("heure"), day ("jour"), dayIntheWeek ("jourSemaine") month ("mois"), and year ("annee"). However, the last column, "nbinterventions" indicates the number of firefighters deployed on that day.

As can be seen from the Table 5.1, some information relates to humidity, air pressure, wind speed, temperature, diseases (e.g., chicken pox), days off, vacations, the number of suicides, the visibility of the moon, and many others. Obviously, not all the information needs to be used in such a study because we are working with time series data based mainly on date/time information and the number of fire calls, and if all attributes are added, it will result in an over-fitted model and will significantly increase the computational and memory requirements. It will also take a lot of time to produce the forecast.

5.4/ Data preparation for Exponential Smoothing techniques

Two datasets are analyzed: the first one (hourly-dataset) contains the number of interventions per hour, while the second one (daily-dataset) contains the average number of interventions per day. Therefore, the hourly dataset consists of 43824 interventions, while the daily dataset carries 1826 interventions starting from "January 1, 2015, 00:00:00" to "December 31, 2019, 23:00:00".

5.4.1/ Outliers Detection

As can be seen in Figure 5.2, and Figure 5.3, there are black dots outside the blue box: these are the outliers.

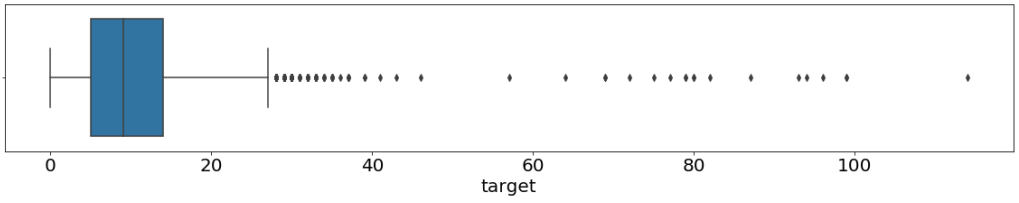


Figure 5.2: Graphical visualization for outliers detection for hourly-dataset

We consider everything above 25 for the hourly-dataset and everything below 6 and above 14 for the daily-dataset as anomalies. The upper and lower bounds were chosen as a

besanconPm25'	'besanconOzone'	hOgnonMontessauxAle'	'humiditeNancy'
debutFinVacances'	'bisonFuteRetour'	hOgnonMontessauxMean'	'luneApparente'
diarrhee_inc100_low'	'diarrhee_inc100'	humiditeBale'	'montbeliardPm10'
diarrhee_inc_up'	'diarrhee_inc_low'	jourAnnee'	'nb_autre'
directionVentNancy'	'directionVentDijon'	montbeliardOxyde'	'nb_suicide'
grippe_inc100'	'ferie'	montbeliardPm25'	'pointRoseeDijon'
grippe_inc_low'	'grippe_inc100_up'	nb_noyade'	'precipitations1hDijon'
hAllanCourcellesMax'	'hAllanCourcellesAle'	phaseLune'	'jour'
hAllanCourcellesStd'	'grippe_inc'	pointRoseeNancy'	'mois'
hDoubsBesanconMean'	'besanconPm10'	precipitations1hNancy'	'nb_feu'
hDoubsVoujeaucourtMax'	'hDoubsBesanconMax'	precipitations3hDijon'	'nebulositeBale'
hDoubsVoujeaucourtStd'	'hDoubsVoujeaucourtAle'	pressionDijon'	'nuit'
hLoueOrnansMean'	'hLoueOrnansMax'	pressionMerNancy'	'pressionBale'
hOgnonBonnalMax'	'hOgnonBonnalAle'	pressionVar3hDijon'	'pressionMerDijon'
'heure'	'hOgnonBonnalStd'	rafalesSur1perDijon'	'pressionVar3hBale'
'hDoubsVoujeaucourtMean'	'pressionMerBale'	'hOgnonMontessauxMax'	'ramadanVeille'
'hLoueOrnansAle'	'pressionNancy'	'hOgnonMontessauxStd'	'temperatureNancy'
'hLoueOrnansStd'	'pressionVar3hNancy'	'humiditeDijon'	'tempsPresentNancy'
'hOgnonBonnalMean'	'rafalesSur1perNancy'	'jourSemaine'	'tendanceBaromNancy'
'diarrhee_inc100_up'	'precipitations1hBale'	'precipitations3hBale'	'precipitations3hNancy'
'rafalesSur1perBale'	'varicelle_inc100_low'	veilleFerie'	'vacances'
ramadanLendemain'	'varicelle_inc_up'	visibiliteNancy'	'directionVentBale'
temperatureDijon'	'visibiliteDijon'	vitesseVentNancy'	'distanceLune'
tempsPresentDijon'	'vitesseVentDijon'	'ramadan'	'grippe_inc100_low'
tendanceBaromDijon'	'besanconOxyde'	'temperatureBale'	'grippe_inc_up'
varicelle_inc'	'bisonFuteDepart'	'tempsPresentBale'	'hAllanCourcellesMean'
varicelle_inc100_up'	'diarrhee_inc'	'tendanceBaromBale'	'hDoubsBesanconAle'
'hDoubsBesanconStd'	'nb_route'	'varicelle_inc100'	'vitesseVentBale'
'montbeliardOzone'	'pointRoseeBale'	'varicelle_inc_low'	'visibiliteBale'
	'nb_accouchement'	'nbInterventions'	

Table 5.1: Attributes of the dataset used in the first contribution

function of the first black dot outside the blue box. The boxplot detects 327 outliers for the hourly-dataset and 39 outliers for the daily-dataset.
To remove what can be considered as anomalies, we replaced the outliers with the lower or upper whisker, and consequently, the datasets have been updated, as can be viewed in Figures 5.4 and 5.5.

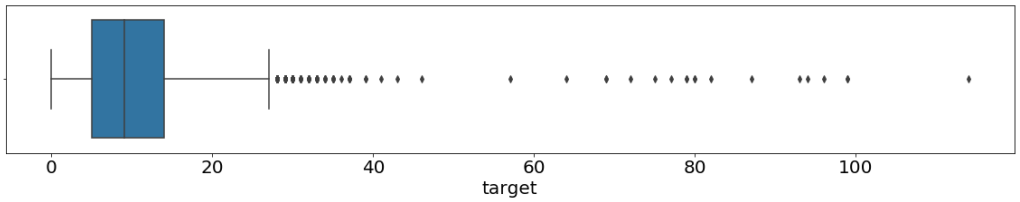


Figure 5.3: Graphical visualization for outliers detection for daily-dataset

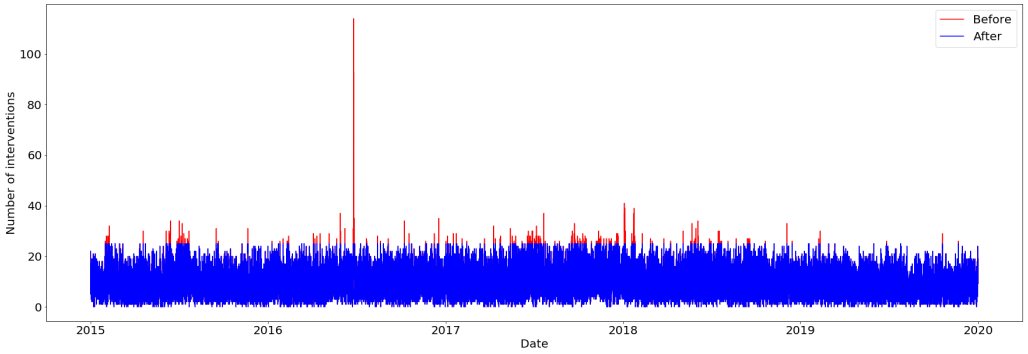


Figure 5.4: Number of firefighters' interventions before and after replacing outliers for hourly-dataset

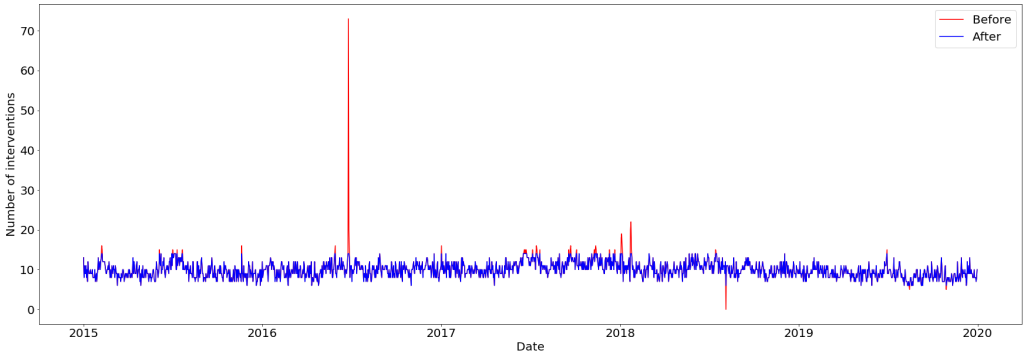


Figure 5.5: Number of firefighters' interventions before and after replacing outliers for daily-dataset

5.4.2/ DATASETS DECOMPOSITION

It is important to perform the decomposition of the datasets to get a structured view of the components used in the Exponential Smoothing methods, such as:

- trend: increasing/decreasing tendencies of firefighters' interventions
- seasonality: repeating cycle
- residual: random variation of the dataset

In this study, both the daily and hourly datasets show an interesting seasonality and trend: the cycle repeats every day/24 hours for the hourly-dataset and every week/7 days

for the daily-dataset, as shown in Figures 5.6 and 5.7. Moreover, the residuals are also reasonable and show different variability over time.

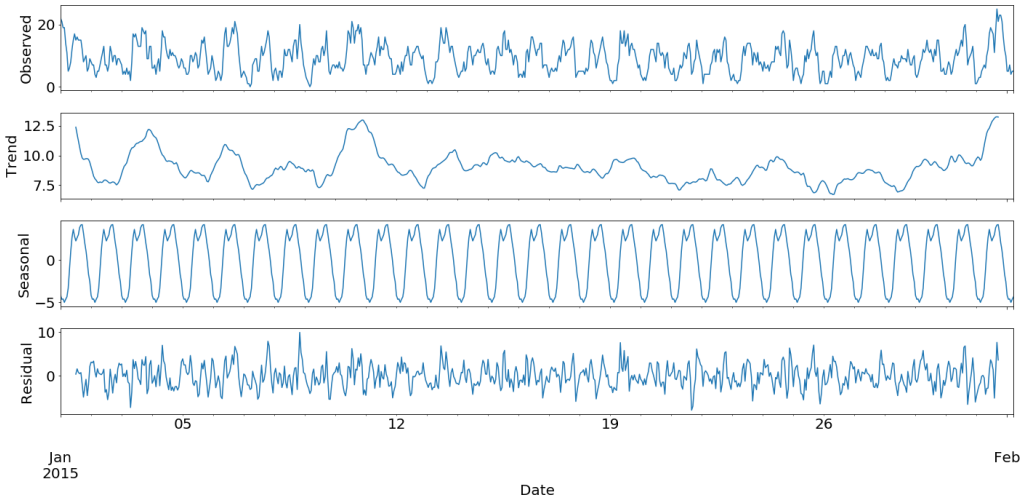


Figure 5.6: Decomposition charts for hourly-dataset

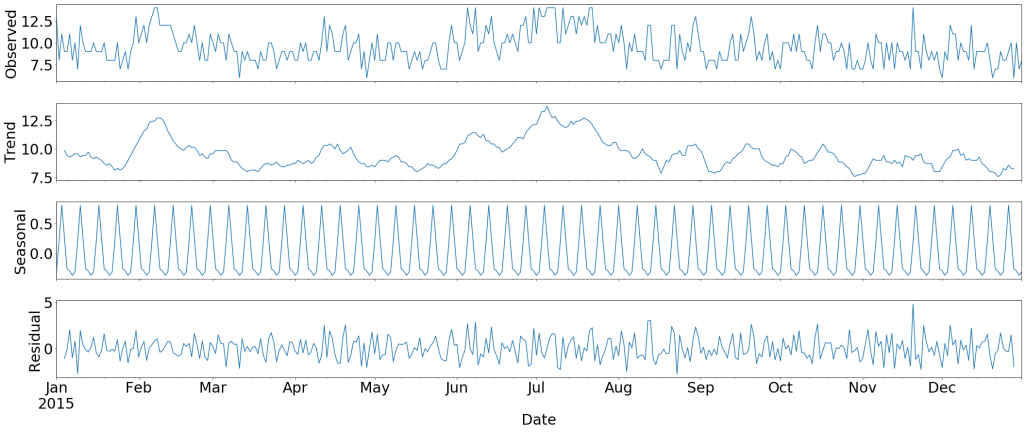


Figure 5.7: Decomposition charts for daily-dataset

Afterwards, each dataset is divided into training and testing, and to back test the three Exponential Smoothing methods, this study uses the walk-forward validation technique by repeating the following steps:

1. training the model using the minimum number of samples in the window
2. prediction for the unique next step
3. evaluation of the predicted value against the real value
4. expansion of the window to include the known value

5.5/ PARAMETERS CHOSEN

5.5.1/ AR, MA AND ARIMA

Auto Regression algorithm does not need any parameters to be chosen or modified as explained in the formula in Section 3.3.1. Nevertheless, different parameters were tested and registered for Moving Average and ARIMA. On the other hand, after trying different values of window sizes for different hours and days (Table 5.2), the best size chosen is 3 hours, which has the minimal values of MAE. On the other hand, to select the values of ARIMA, the following parameters should be taken into consideration:

1. p: Autocorrelation (ACF) is the calculated coefficient indicating how much the values in the same time series are similar, meaning how much the values are correlated, and it ranges from -1 to 1. A coefficient of zero means that there is no relationship between the data points. The ACF chart shows the correlation value as a function of a specific point in time, called lag, with the blue bars representing the error ranges.

   The order of AR term was basically taken to be equal to the number of lags that crosses the significance limit in the Autocorrelation (Figure 5.8). It is observed that the ACF lag 4 is quite significant. Then, p was fixed to 4.

2. d: let us use the Augmented Dickey Fuller (ADF) test to see if the number of interventions is stationary. ADF is a statistically significant test that shows whether a time series dataset is stationary or not by specifying the p_value. The coefficient found is $5.12e^{-28}$, which is lower than the significant level of 0.05. This means that no differencing is needed. Let's fix d to 0.

3. q: one lag above the significance level was found, thus q=1 (Figure 5.8).

Figure 5.9 shows the actual number of interventions versus the predicted ones after applying a moving average transformation overlaid by 3 hours.

Table 5.2: Different window sizes for Moving Average Algorithm.

Window Size	MAE
1 hour	1.477
2 hours	1.360
3 hours	1.349
4 hours	1.359
5 hours	1.387
10 hours	1.541

Hence, after determining the values of p, d, and q, ARIMA model is fitted by using order (4,0,1).

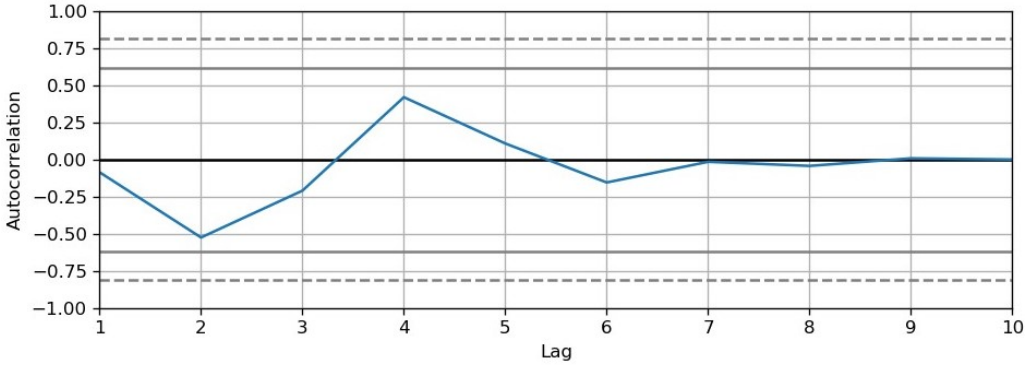


Figure 5.8: Autocorrelation plot

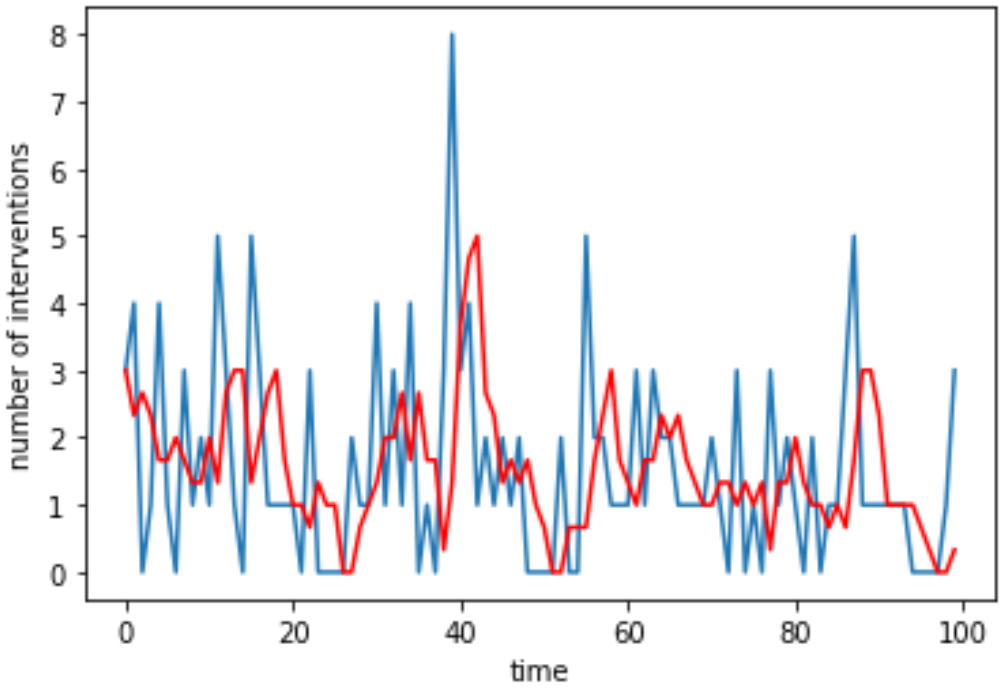


Figure 5.9: Actual versus predicted number of interventions.

5.5.2/ EXPONENTIAL SMOOTHING

The sensitivity of the predictions depends on the smoothing constants. Larger values of alpha (α) form a forecast that is more sensitive to recent observations, while in contrast, smaller values give a dampening result. The same concept applies to beta (β), which emphasizes recent trends over older observed values.

In this chapter, the smoothing parameters were selected as a function of the minimum

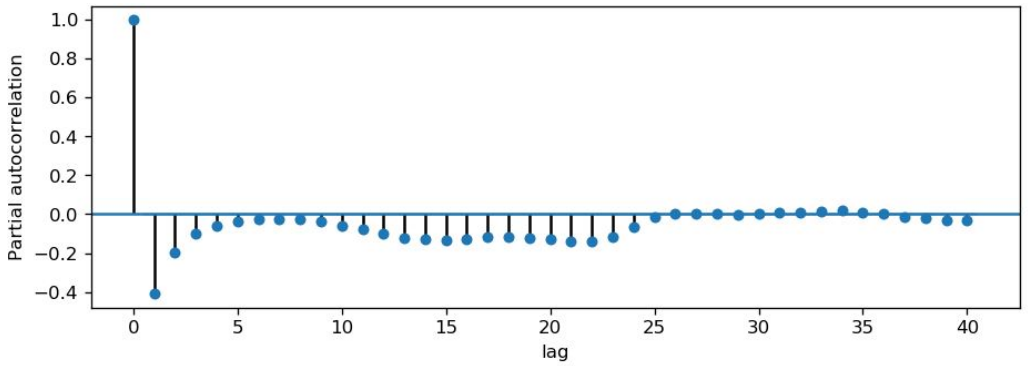


Figure 5.10: Partial autocorrelation plot

values of mean absolute error (MAE) and root mean squared error (RMSE) so that the forecasts are more accurate. Different values of α and β were tried on the datasets. The concept is to repeat the loop 99 times for simple Exponential Smoothing, whereas for Holt's method the loop is repeated 99*99 times because it has two smoothing parameters. α and β are ranged from 0 to 1 and the loops increase the value of α and β by 0.01 at each iteration. On the other hand, the seasonality used for the Holt-Winters' method is picked up depending on the seasonal curve presented in section 5.4.2.

The obtained optimal constants for the hourly-dataset are alpha = 0.9, beta = 0.05 (Table 5.3) and the seasonality is 24 hours. On the other side, the optimal values for the daily-dataset are alpha=0.1, beta=0.05 (Table 5.4), and the seasonality is 7 days.

Table 5.3: The RMSE measures using different value of smoothing constant (α)

Daily-dataset		Hourly-dataset	
alpha	RMSE	alpha	RMSE
0.1	1.409	0.1	4.775
0.5	1.464	0.5	3.476
0.9	1.624	0.9	3.172

Table 5.4: The RMSE measures using different value of smoothing constants (α) and (β)

Daily-dataset			Hourly-dataset		
alpha	beta	RMSE	alpha	beta	RMSE
0.1	0.05	1.408	0.9	0.05	3.175
0.5	0.05	1.451	0.9	0.5	3.246
0.1	0.5	1.656	0.4	0.5	3.961

5.5.3/ BUILDING DATA WITH PROPHET

For the prophet's preparation, a new dataframe should be found: a new column is added to the data that emerges from years, months, days, and hours. Then, this column is renamed to 'ds' and the predicted output presented in the data under the name of nbinterventions is renamed as 'y' as shown in Table 5.5.

Figure 5.11 helps to visualize the forecast of the dataframe where the black dots display actual measurements, the blue line indicates Prophet's forecast, and the light blue shaded line shows uncertainty intervals. It appears that interventions have increased over the years slightly.

Table 5.5: New dataframe for the dataset.

Index	y	ds
0	1	2006-01-01 00:00:00
1	1	2006-01-01 01:00:00
2	0	2006-01-01 02:00:00
3	2	2006-01-01 03:00:00
4	4	2006-01-01 04:00:00

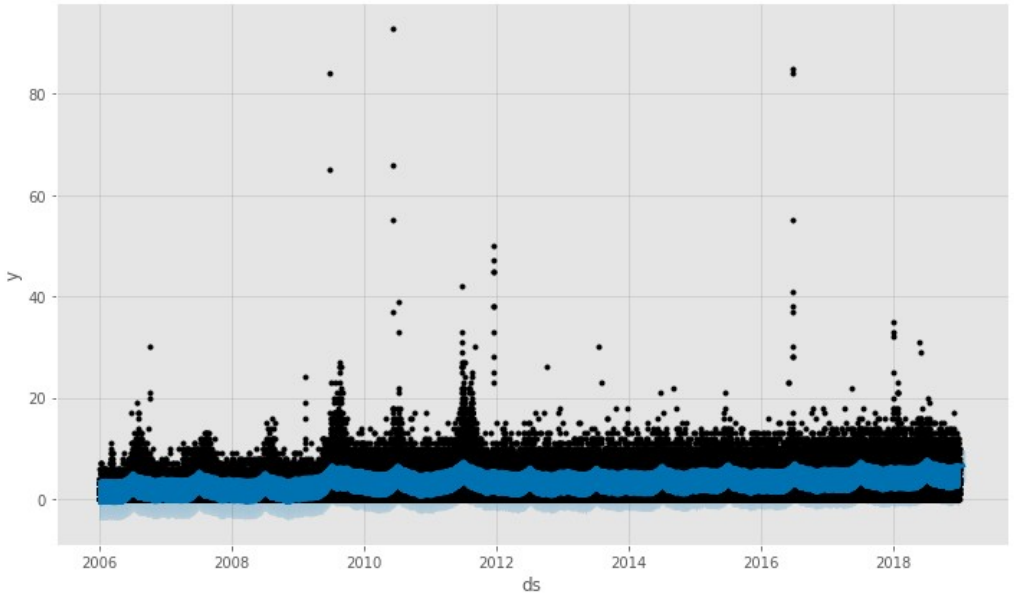


Figure 5.11: Forecast for Prophet algorithm

5.6/ RESULTS AND DISCUSSION

Considering having 24 hours per day, 7 days per week, 30 days per month, and 365 days per year, three algorithms have been implemented: AR, MA and ARIMA on the dataset in addition to the prophet. Statistical features of firemen predictions for different time intervals have been registered in Table 5.6 and for every year from 2006 until 2018 in Table 5.7. However, Figures 5.12, 5.13, 5.14 show statistical features for AR, MA and ARIMA respectively. Moreover Tables 5.8 and 5.9 present RMSE and MAE for different prediction models for hourly and daily datasets over various time periods.

On the other hand, let us overview the result of the forecast by illustrating a breakdown of the former elements (Figures 5.15, 5.16) for daily, weekly, and yearly trends using the prophet tool. The number of interventions during a trimmed time slot is shown in Figure 5.17 where:

- yellow plot represents the actual number of interventions y
- purple plot indicates the prediction yhat
- blue and red plots show the upper and lower bound of prediction respectively

Table 5.6: Statistical features using AutoRegression for different time slots.

Time	MAE	MSE	RMSE
1 hour	0.307	0.094	0.307
2 hours	0.403	0.171	0.414
12 hours	1.205	1.932	1.390
1 day	1.288	2.128	1.459
5 days	1.168	1.822	1.350
1 week	1.209	2.057	1.434
1 month	1.213	2.123	1.457

Mean Absolute Error for AR, MA, ARIMA, and prophet are compared in Figure 5.18. The best algorithm in terms of fewer errors in most cases is ARIMA represented by the gray line. Among most of the years, starting from 2006 until 2018, this gray line has reached the minimum mean absolute error and root mean squared error, compared to autoregression, moving average algorithms and the prophet tool. Thus, the second, third, and fourth places are reserved respectively for Moving Average, Autoregressive, and Prophet.

Generally, since ARIMA(p,d,q) stands for Auto Regressive Integrated Moving Average, it is logical to conclude that it combines AR (parameter p) and MA (parameter q) models.

Table 5.7: Statistical features using AR, MA ,ARIMA, and prophet from 2006 until 2017.

	AR		MA		ARIMA		Prophet	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
2006	1.481	2.046	1.349	1.86	1.018	1.307	2.55	3.53
2007	1.601	2.064	1.429	1.924	1.376	1.822	1.95	2.97
2008	1.496	1.952	1.385	1.868	1.263	1.644	1.31	1.63
2009	2.374	3.35	1.854	2.787	1.414	1.904	3.25	5.69
2010	2.161	3.058	1.847	2.716	1.629	2.154	2.00	2.23
2011	2.574	3.676	2.09	2.922	1.699	2.247	6.00	11.00
2012	1.99	2.5	1.84	2.415	1.642	2.031	2.44	2.87
2013	1.972	2.478	1.83	2.392	1.682	2.14	2.33	2.66
2014	2.04	2.545	1.874	2.451	1.504	1.843	2.44	2.90
2015	2.145	2.678	1.939	2.525	1.829	2.43	2.73	3.15
2016	2.223	3.137	2.026	2.807	1.898	2.31	2.45	2.99
2017	2.359	2.917	2.111	2.738	1.941	2.544	2.86	3.49
2018	2.552	3.216	2.24	2.929	2.075	2.742	1.85	2.60

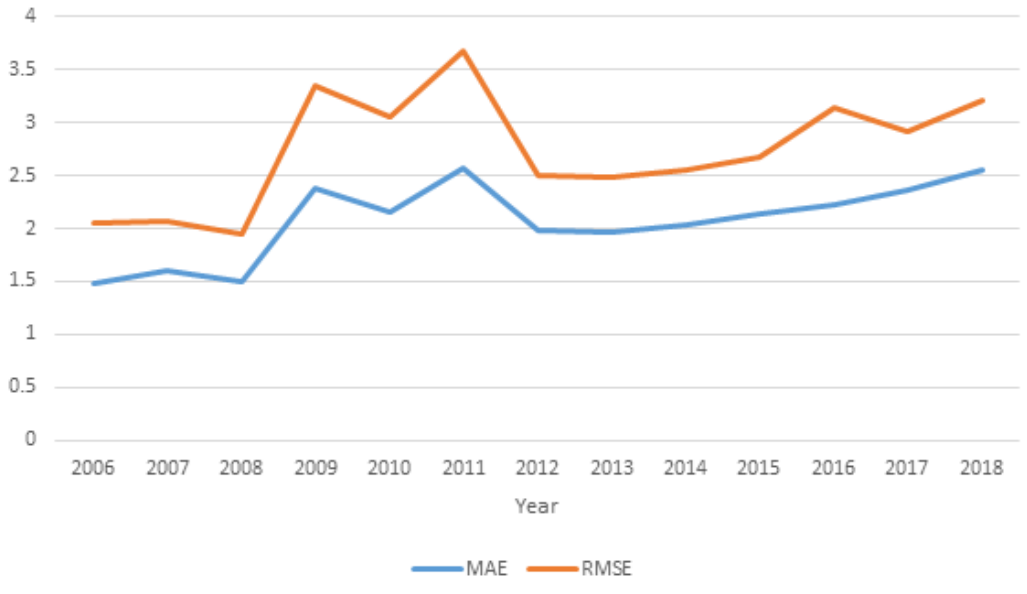


Figure 5.12: Statistical features over many years using AR

Other than that, ARIMA ensures the stationarity of the model (parameter d), unlike AR and MA. Therefore, by applying the components of these two models together, the probability of making errors will be reduced, as shown in the experiments. It is important to

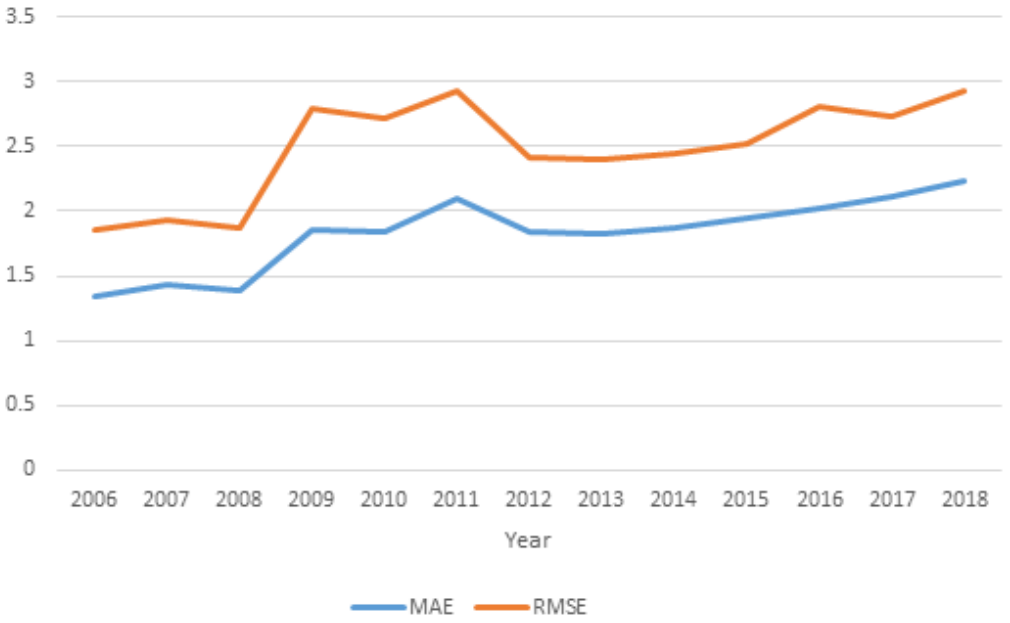


Figure 5.13: Statistical features over many years using MA

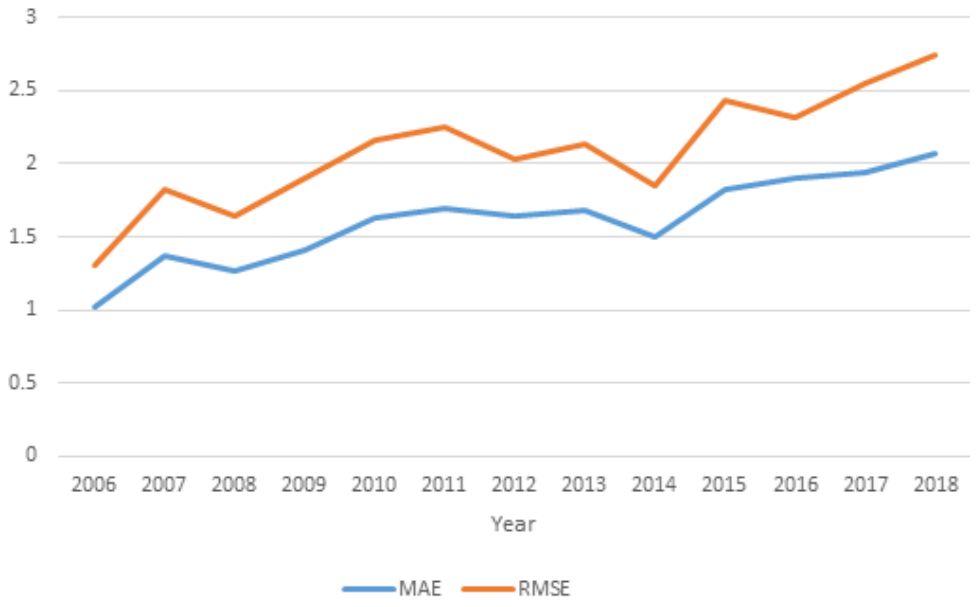


Figure 5.14: Statistical features over many years using ARIMA

mention that ARIMA is more complex than applied algorithms since it requires more time to identify the excessive number of parameters p, d and q. In contrast, when comparing eXtreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM) algorithms applied and experimented in [Cerna et al., 2020] together with ARIMA from 2006 to 2014, it seems that ARIMA has the lowest root mean squared error values. However, XGBoost has the minimum RMSE values for 2015, 2016 and 2017. This result reflects that XGBoost is better for long-term forecast usage.

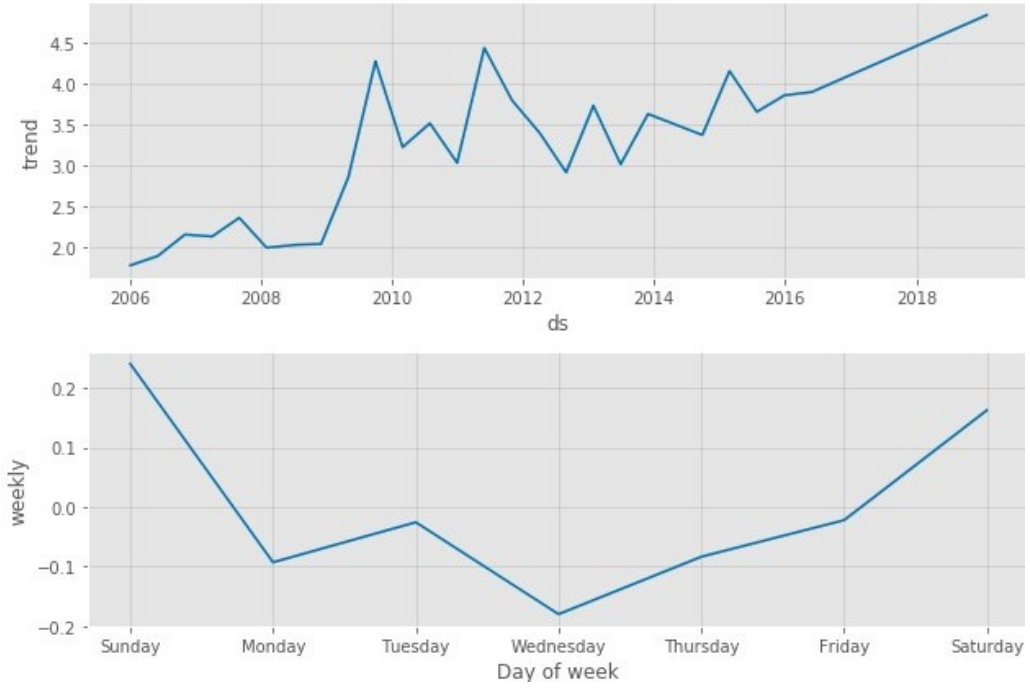


Figure 5.15: Trend and weekly prediction

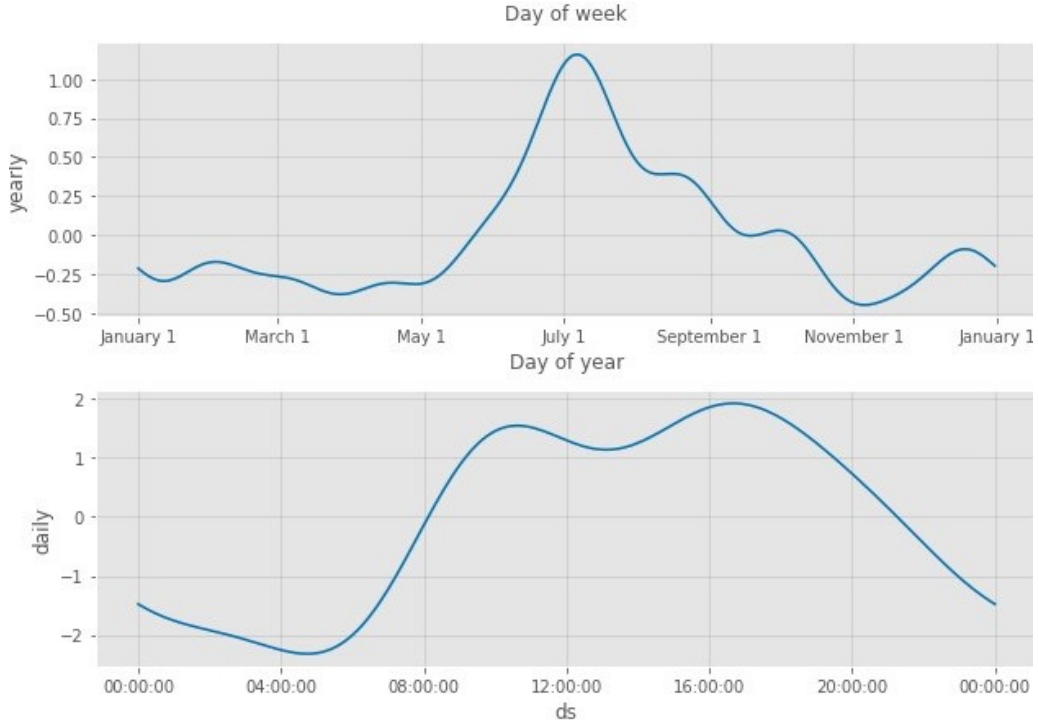


Figure 5.16: Yearly and daily prediction

On the other side, by analyzing the prophet result, it is very clear that the number of interventions by firefighters increases highly during the weekend (Saturday and Sunday) and reaches its minimum during the middle of the week (Wednesday). This interpretation

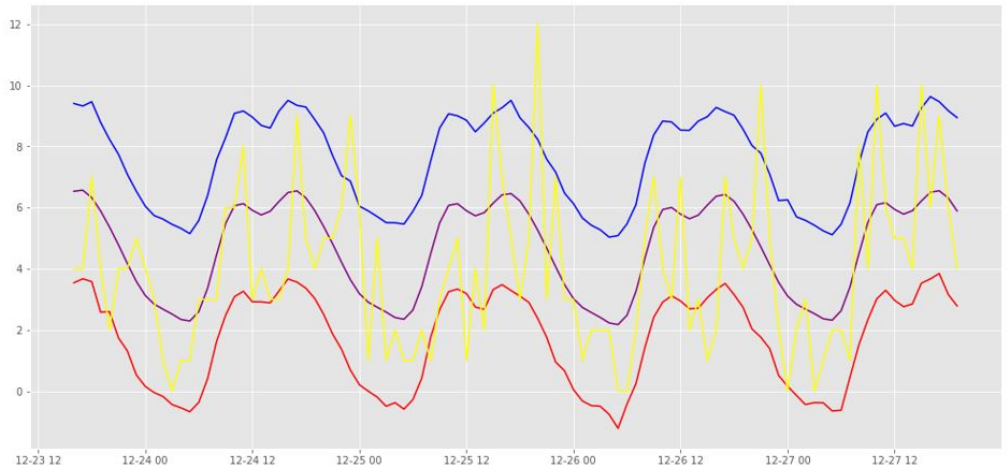


Figure 5.17: Number of firefighters' interventions using Prophet

corresponds to official days off in France. Also, the number of interventions increments slightly starting in the month of May and reaches its maximum in July, then decreases gradually till November. The fluctuation of interventions per month reflects that during the summer, incidents are more likely to happen than in other seasons. On the other hand, the daily seasonality illustrates that the number of interventions increases during the morning starting around 5:00 am and reaches higher values between 11:00 am and 5:00 pm. It's very logical to link this variation with the departure and return time from work.

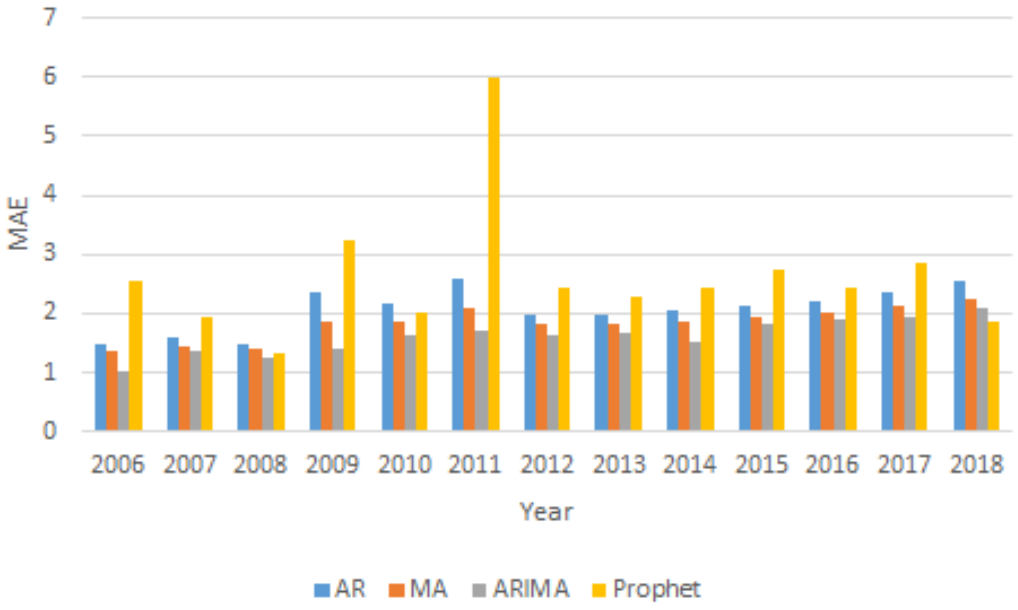


Figure 5.18: Mean absotule error comparison for AR, MA and ARIMA

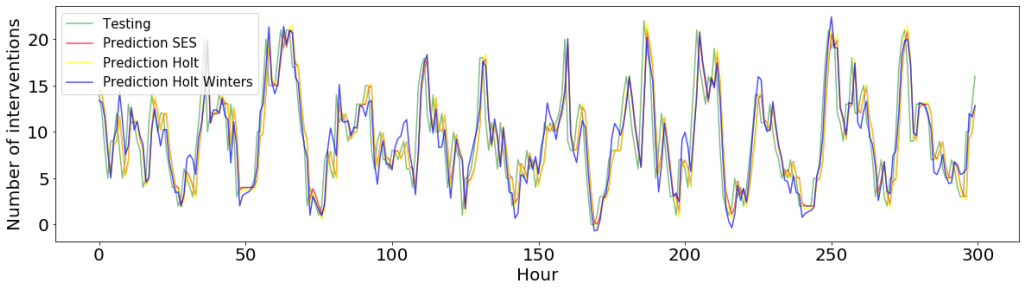
To check the efficiency of the models used, forecasting was conducted for different time intervals in the future, and to compare the obtained results with a reference, the persis-

tence model is considered as the baseline, and MAE and RMSE are calculated. Moreover, the prediction of the number of firemen for 300 hours in 2019 for the hourly-dataset is shown in Figure 5.19 and the result of the prediction for the daily-dataset for the whole year 2019 is displayed in Figure 5.20.

Period	SES		Holt		Holt-Winters		Persistence	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
1 h	0.47	0.47	0.605	0.605	0.631	0.631	1	1
2 h	1.485	1.259	1.55	1.356	0.947	0.906	1.581	1.5
3 h	2.821	2.241	2.73	2.3	2.15	1.762	2.646	2.333
5 h	3.072	2.76	3.088	2.801	2.727	2.401	3.033	2.8
7 h	2.842	2.458	2.851	2.481	2.394	1.983	2.803	2.429
12 h	3.218	2.79	3.237	2.826	3.44	2.76	3.215	2.833
1 day	3.05	2.543	3.062	2.566	3.17	2.525	3.021	2.542
2 days	3.339	2.616	3.348	2.63	3.347	2.488	3.189	2.5
3 days	3.251	2.476	3.251	2.481	3.1	2.383	3.238	2.458
4 days	3.172	2.362	3.175	2.368	3.06	2.322	3.135	2.326
5 days	3.136	2.367	3.138	2.369	3.046	2.33	3.129	2.342
10 days	3.237	2.514	3.312	2.575	3.149	2.46	3.206	2.471
15 days	3.157	2.425	3.23	2.484	2.991	2.336	3.121	2.376
1 month	3.223	2.499	3.298	2.564	2.957	2.306	3.179	2.455
2 months	3.219	2.481	3.293	2.542	2.926	2.23	3.179	2.428

Table 5.8: RMSE and MAE for different prediction models for hourly-dataset over various time period

Figure 5.19: Various models to predict the number of firefighters' interventions during 300 hours in January 2019 for hourly-dataset

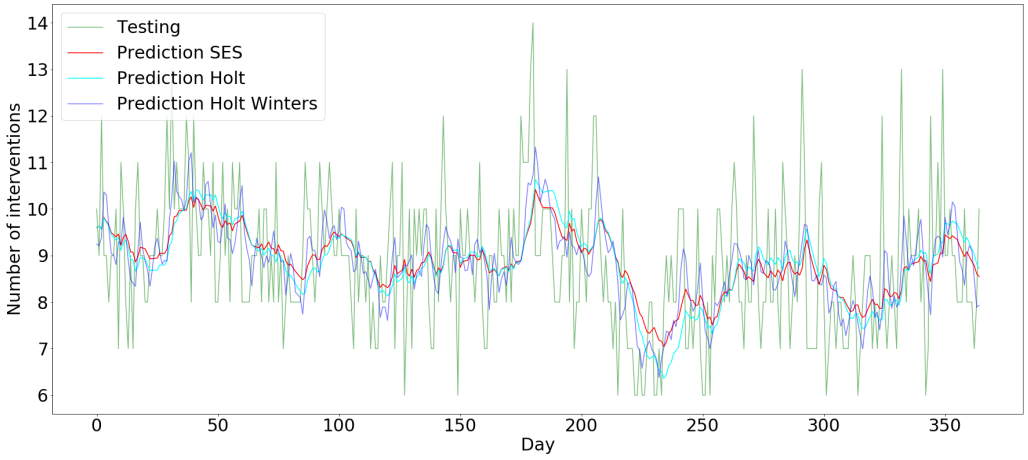


To show the results for multiple periods and since seasonality is evident in this study for both the daily and hourly datasets, we calculated the average number of firefighters'

Period	SES		Holt		Holt-Winters		Persistence	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
1 day	0.402	0.402	0.414	0.414	0.75	0.75	1	1
2 days	0.533	0.52	0.531	0.52	0.55	0.479	1	1
3 days	1.467	1.155	1.473	1.16	1.571	1.189	1.915	1.667
5 days	1.238	1.004	1.241	1.005	1.481	1.247	2	1.6
7 days	1.234	1.025	1.236	1.026	1.394	1.128	1.773	1.429
2 week	1.421	1.203	1.42	1.202	1.442	1.191	1.964	1.571
4 week	1.218	0.929	1.218	0.928	1.201	0.95	1.69	1.289
8 week	1.323	1.017	1.322	1.016	1.304	1.03	1.778	1.411
16 weeks	1.204	0.946	1.203	0.945	1.197	0.94	1.573	1.205
32 weeks	1.357	1.075	1.356	1.075	1.332	1.07	1.654	1.246
48 weeks	1.405	1.117	1.404	1.118	1.353	1.09	1.659	1.265
1 year	1.409	1.114	1.408	1.114	1.37	1.093	1.682	1.274

Table 5.9: RMSE and MAE for different prediction models for daily-dataset over various time period

Figure 5.20: Various models to predict the number of firefighters' interventions during the whole year 2019 for daily-dataset



interventions on an hourly basis from "00:00:00" to "23:00:00" for the hourly-dataset and on a weekly basis from Monday to Sunday for the daily-dataset. Subsequently, the same concept is processed by applying Simple Exponential Smoothing, Holt's, Holt-Winters', and persistence models. The results for the different methods used are illustrated in Figure 5.21 and Figure 5.22.

From Figure 5.23 and Figure 5.24, it can be seen that the value of RMSE for hourly-

Figure 5.21: The average number of firefighters' interventions over the hours of a day for hourly-dataset

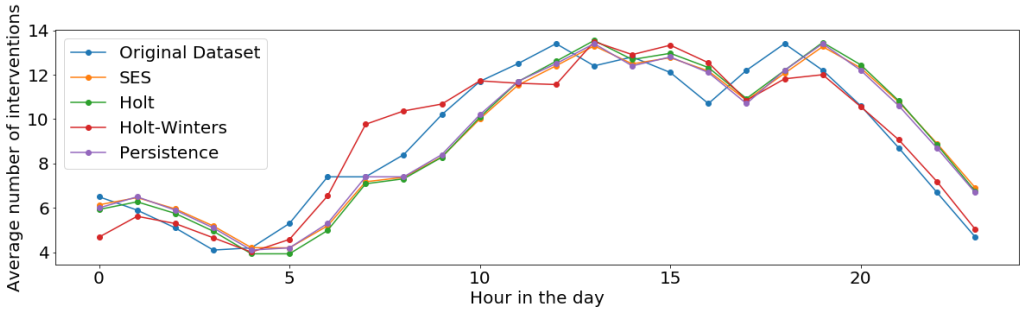
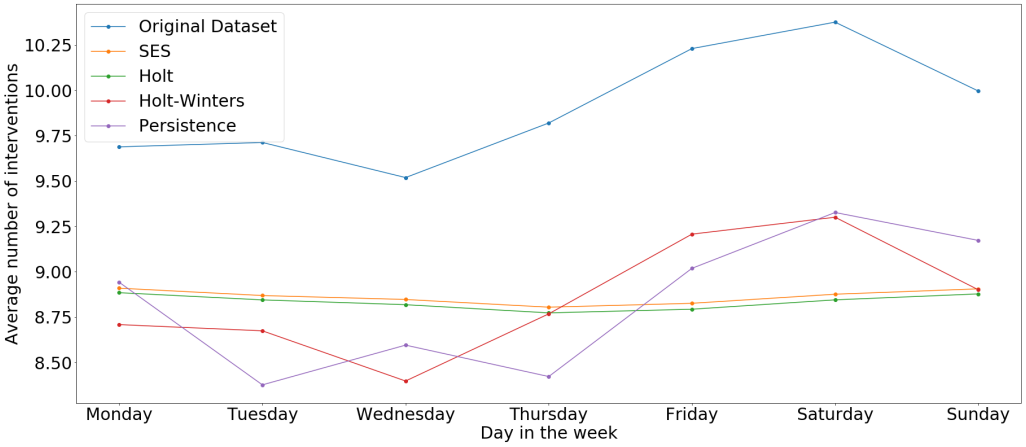


Figure 5.22: The average number of firefighters' interventions over the days of the week for daily-dataset



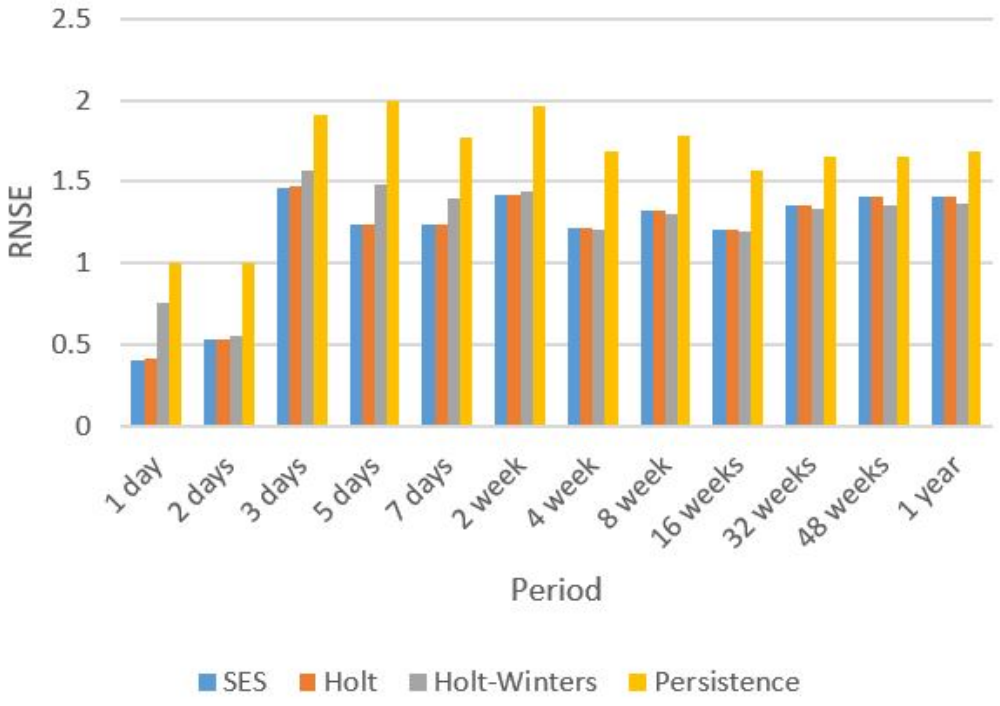
dataset is increased with the increase of alpha. However, the opposite is observed for the daily-dataset as the RMSE decreases with the increase of alpha. The selected optimal values resulted from the minimal RMSE. These results reflect that Exponential Smoothing assigns larger weights to the recent observations in the hourly-dataset and fewer weights in the daily-dataset. This means that α has a smaller effect and gives more importance to the recent observations in the hourly-dataset, while the data in the daily-dataset is less sensitive to the recent changes. Furthermore, the optimal value of beta is 0.05, which is very close to zero. This means that more weight has been given to the past trends in the estimation of current trends.

After selecting the smoothing constants that produce less error, MAE and RMSE are calculated for each prediction period for the hourly and daily datasets to measure the analysis performed during the forecasting process. The use of single and double Exponential Smoothing in this work is not effective compared to the Holt-Winters' method, which gives the lowest prediction error over time. In other words, as can be seen in Figure 5.23 and Figure 5.24, when Holt-Winters' method is used, the RMSE decreases as the prediction period increases. This result reveals that Triple Exponential Smoothing is a

feasible technique used in this research because both the daily and hourly datasets are seasonal.
Finally, the compilation of the average number of firefighters' operations during the days of the week for the hourly-dataset and during each hour of the day for the daily-dataset reveals many relevant facts. Figure 5.21 and Figure 5.22 show that Holt-Winters' method has the most accurate values of prediction compared to the original values of interventions. Additionally, it is observed in Figure 5.21 that the firemen services increase from 5:00 am, tend to peak, and remain broadly stable throughout the day before gradually decreasing at 7:00 pm. This is very reasonable because the number of vehicles and flow of people out of their homes is higher during the day, causing rush hours and thus more risk of damage and incidents.

On the other hand, it can be seen in Figure 5.22 that the firefighters have the highest number of interventions from Friday to Sunday, with fewer fluctuations on weekdays. Unsurprisingly, weekends are the riskiest days for fatal crashes.

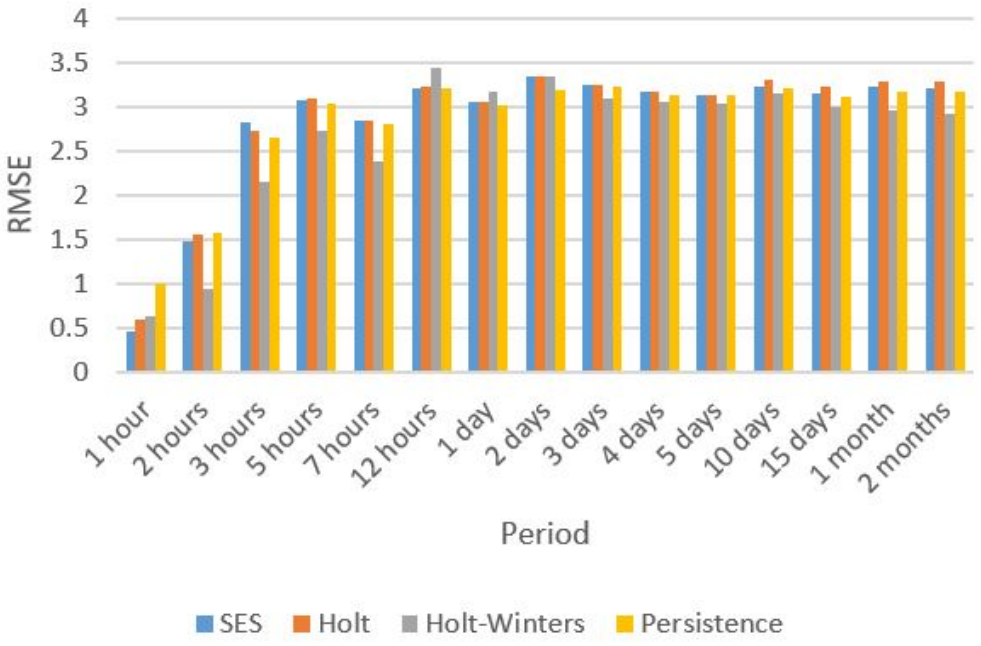
Figure 5.23: Prediction results on hourly-dataset



5.7/ CONCLUSION

It is very evident that the prediction of the number of firefighters' interventions is not a simple random process but is influenced by hourly and weekly changes related to human activities.

Figure 5.24: Prediction results on daily-dataset



Autoregression, Moving Average and Auto Regressive Integrated Moving Average have been implemented, as well as a Facebook tool for time series forecasting called Prophet. For each model, statistical parameters have been calculated and compared between each other and then compared between other results previously done. In general, as many researchers agreed, no hypothesis or rule selects a better algorithm in all-time series forecasting. The choice of the technique used depends on the specific prediction problem, taking into account trends, seasonality, variables, size of the dataset, etc. Statistical metrics indicate that ARIMA is the best model compared to AR and MA as it combines first the characteristics of these two algorithms and second the stationarity of the model. In contrast, XGBoost fits better than ARIMA for long-term prediction.

On the other hand, Exponential Smoothing is a prominent tool to use in such a study that provides reliable forecasting. Statistical characteristics and graphical exploration of data have been presented to find the best Exponential Smoothing technique, and three models were developed and then compared to each other and the baseline. Based on the prediction error, the least variation measures were found for the method Holt-Winters' since it takes into account seasonality, which is the main component of hourly and daily datasets.

6

ANOMALIES AND BREAKPOINT DETECTION DURING COVID-19

In the last three years, COVID-19 has been the focus of most research in various fields. Handling big data with worldwide information on COVID-19 cases and deaths has been the main task of machine learning. In this chapter, the focus is on predicting the number of firefighters' missions during the sensitive phase of the global pandemic COVID-19. Experiments applied to a dataset from 2016 to 2021 provided by the fire and rescue department SDIS 25 in the Doubs region of France have shown accurate predictions and revealed the presence of a turning point in August 2020 due to an increase in coronavirus cases in France.

6.1/ INTRODUCTION

The World Health Organization (WHO) recorded the first pneumonia of an unknown cause on December 31, 2019, in Wuhan, China. A few days later, on January 24, 2020, it was confirmed that the virus had reached France. Countries could never have imagined that such a devastating viral disease would emerge and a pandemic could cause morbidity around the world. The ability of public health actors, populations, and institutions to prepare for and respond effectively to such a crisis in order to maintain essential functions was not satisfied worldwide. Hospitals faced bed shortages, delayed non-urgent surgeries, and mobilised all medical staff resources. This inevitably made the fire and rescue services (SIS) in France the first actors of aid and emergency care providers in the operational response to this epidemic. They are the key players in this crisis, in caring for people and in supporting the health system by participating in transports between hospitals by land or by helicopter. They also run facilities for the elderly ('Etablissement d'Hébergement Pour Personnes Agées Dépendantes EHPAD'): examination, disinfection, distribution of meals, etc. Therefore, the use of machine learning is very important in such a case: predicting the number of firemen interventions can directly lead to a reduction in

financial, material, and human resources. Consequently, the efficiency of emergency operations during this pandemic will be improved.

The database for this study, illustrated in Figure 6.1 was provided by the Fire and Rescue Department, SDIS 25, in the Doubs-France region. It contains hourly recorded operations from January 2015 to June 2021. The aim of this chapter is to predict the number of firefighters by performing several steps, starting with the reduction of the number of features, the detection of the breakpoint related to the coronavirus period, the comparison of the selected features before and after the COVID-19, and finally the detection of the anomalies.

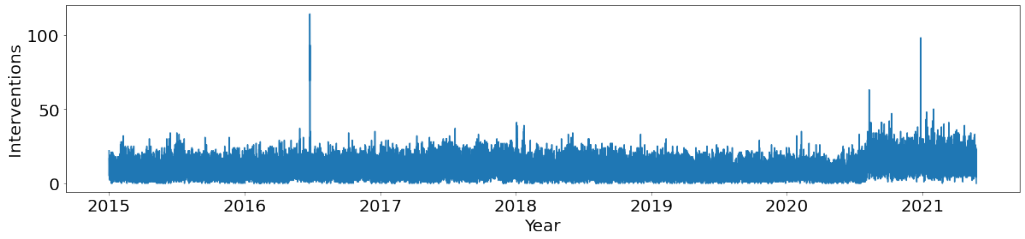


Figure 6.1: Dataset presentation

This chapter presents the related works to this contribution, presents the feature selection method used to reduce the number of attributes in the database, and describes how the breakpoint was selected and its relationship to the selected features. Also, anomaly detection is presented based on experiments and results, and finally a brief conclusion is drawn.

6.2/ STATE OF THE ART

Machine learning and artificial intelligence have played an important role in research, healthcare, and even agriculture during the COVID-19 period. ML technology allows computers to interpret large amounts of data to quickly identify patterns and insights and understand the pathobiology of this disease. Numerous studies have been proposed in various countries to predict the spread of COVID-19. In India, a study was conducted using linear regression, multilayer perceptron, and vector autoregression to predict the spread of this disease [Sujath et al., 2020]. Likewise, a hybrid machine learning approach based on an adaptive network-based fuzzy inference system and a multilayer perceptron imperialist-competitive algorithm was proposed to predict the individuals infected by coronavirus and mortality in Hungary [Pinter et al., 2020].

Furthermore, Machine Learning helped detect COVID-19 at an early stage, which helped monitor disease progression and potentially reduce mortality. One study was validated on a dataset of chest X-rays and CT scans compared to popular deep learning-based

feature extraction frameworks for automatic COVID-19 classification by sorting subjects as control or COVID-19 [Kassania et al., 2021].

In addition, Artificial Intelligence played an important role in combating the growing pandemic. If enough data is trained by a deep learning model, it can help in finding an effective vaccine candidate by detecting patterns in the data [Keshavarzi Arshadi et al., 2020]. A research paper proposed a silico approach for the prediction and design of a multi-epitope vaccine (DeepVacPred), which helps speed up the process of vaccine development. The applied framework predicts 26 vaccine subunits from the existing SARS-CoV-2 spike protein sequence and has proven to cope with recent mutations of the virus [Zhang et al., 2020].

6.3/ METHODOLOGY

In this section, the methodology used for predicting the number of firefighters' interventions is explained. First, the number of attributes in the dataset was reduced by applying a feature selection technique, then breakpoints were detected, and finally anomalies were found and replaced. At each step, statistical features, specifically mean absolute error and root mean square error, were calculated to evaluate the method used. However, since the choice of hyperparameters directly reflects the performance of the model, Optuna [Akiba et al., 2019], a hyperparameter optimization system, was used to optimize the parameters of the XGboost algorithm [Chen et al., 2015], specifically learning_rate, max_depth, random_state, n_estimators, and n_jobs.

6.3.1/ DATASET

As mentioned earlier, the dataset used in this chapter is identical to the previous one, with the scope reduced to 2015 to 2020. In addition, three new attributes related to coronavirus were added to the dataset, making this study possible. Figures 6.2, 6.3, 6.4, 6.5 and 6.6 provide an overview of the entire dataset by showing the number of interventions over different years, one month, one year, one week and one day respectively. These graphs show a very logical analysis of the number of interventions demonstrated in the previous chapter in relation to firefighter characteristics and whose interventions are related to time of day, day of week, months, and years. As can also be seen, Figure 6.2 shows a sharp increase in the number of interventions from 2020 with the occurrence of the coronavirus pandemic.

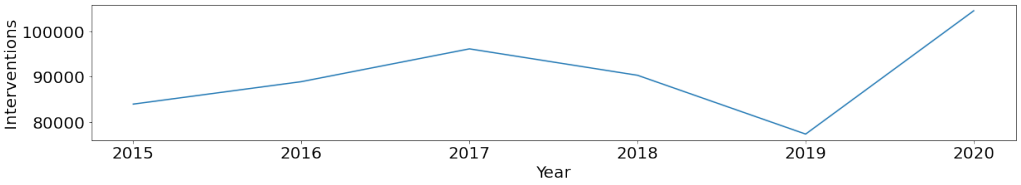


Figure 6.2: Interventions over the years

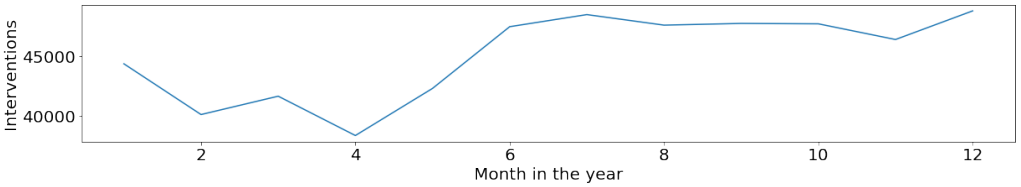


Figure 6.3: Interventions over 12 months

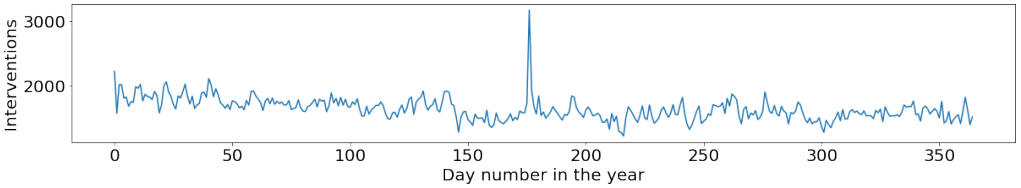


Figure 6.4: Interventions during 365 days

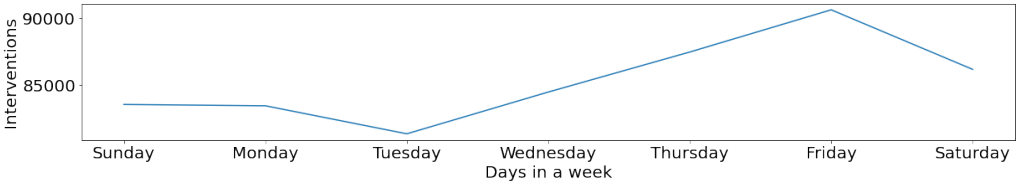


Figure 6.5: Interventions during one week

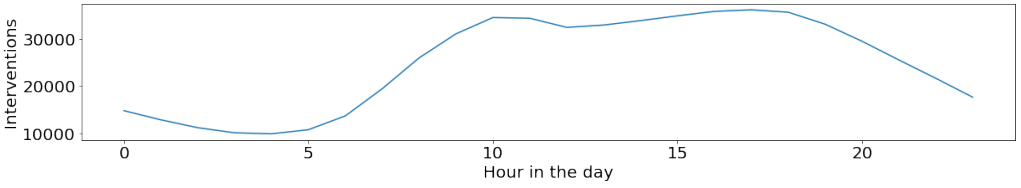


Figure 6.6: Interventions during 24 hours

6.3.2/ Feature Selection

The presence of 1573 attributes certainly complicates the processing of the dataset. Therefore, feature selection is required before we proceed to the next stage of this study. The idea is to retain only useful features by removing irrelevant attributes, since adding more and more variables to a dataset increases the overall complexity and could reduce the accuracy of a classifier. Thus, a feature is considered irrelevant if it has been removed without affecting the performance of the model. The feature selection approach used in

this chapter is called “feature importance,” which is a built-in class where a score is assigned to each feature in the dataset, and the higher the score, the more relevant the feature is [Zien et al., 2009].

Three machine learning algorithms were implemented: XGBOOST [Chen et al., 2015], Extra Tree Classifier [Geurts et al., 2006] and Random Forest Regressor [Segal, 2004]. For each algorithm, feature importance was calculated and the common best top 16 attributes were extracted, including the ’lockdown’ feature corresponding to the period of COVID-19.

6.3.3/ BREAKPOINT DETECTION

In the dataset used (Figure 6.1), it is quite evident to recognize an increase in the number of firemen interventions starting in August 2020, which is closely related to the augmentation in the number of COVID-19 cases in France, as can be seen in Figure 6.7, which shows the number of active cases since the emergence of coronavirus disease in France until today according to Worldometer statistics. Thus, the objective of change point detection is to highlight the direct impact of coronavirus on the number of firefighters’ interventions. For this purpose, a Python library for the analysis of breakpoints called ’rupture’ was applied [Truong et al., 2020].

After running the code, a change point was detected on August 5, 2020, and is presented in a green dotted line in Figure 6.8. Therefore, we divided the dataset into two periods separated by the breakpoint: pre and post COVID-19 spread.

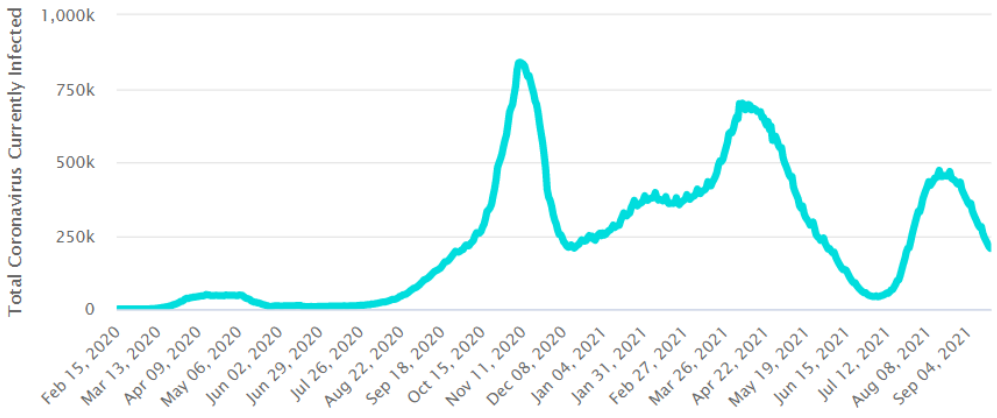


Figure 6.7: Breakpoint detection

Further action was taken by rechecking the order of feature importance to ensure that the attributes related to the COVID-19 differed if they were before or after the breakpoint found. Table 6.1 shows that the score of the three features ’confinement1’, ’confinement2’, and ’couvrefeux’ (representing ’lockdown1’, ’lockdown2’, and ’curfew’, respectively) associated with coronavirus disease changes immediately before and after the

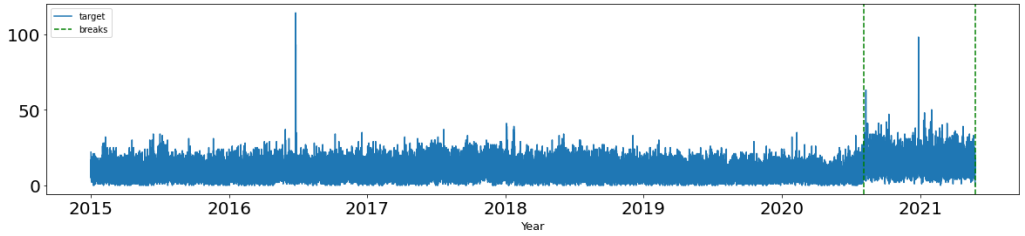


Figure 6.8: Breakpoint detection

identified breakpoint. Indeed, the French president announced a nationwide lockdown starting March 17, 2020 until May 10, 2020. A second lockdown followed on October 30, 2020, and lasted until mid-December. The curfew, however, will remain in effect from December 15, 2020 to June 20, 2021.While the goals of the lockdown and curfew may seem similar, there is a major difference between the two: during a curfew, people are forced to stay at home for a specific number of hours, and major services remain closed for a certain period of time.

Period	Before 5/8/2020	After 5/8/2020
Feature	Feature importance score	
confinement1	0.024589	0.000000
confinement 2	0.000000	0.018046
couvrefeux	0.000000	0.056813

Table 6.1: Feature Importance Before and After COVID-19 peak period

6.3.4/ ANOMALIES DETECTION

Anomaly detection is required in such huge metadata to find elements that trigger suspicions that might be different from the majority of the data. Logically, these anomalies can be associated with rare events such as water floods, large fires, power outages, and, of course, COVID-19 disease. In this work, Isolation Forest, an advanced statistical learning technique, was used by importing the PyCaret library [Ali, 2020]. This method assumes that some data points are more dominant than others, making outliers susceptible to the isolation mechanism. Anomaly detection was done by selecting a random subset of data. Then the threshold value is assigned (lower and upper threshold): any value in the minimum and maximum range of the selected feature.

After that, if the value of the data point is less than the threshold, the selected data point moves to the left branch. The steps are repeated until the maximum depth is reached or isolation is assigned. The anomalies are represented by the value -1, while normal points are represented by 1.

1122 anomalous data points are presented as red dots in Figure 6.9. It can be clearly seen that most of the outliers are detected in the time after the breakpoint found in Section 6.3.3.

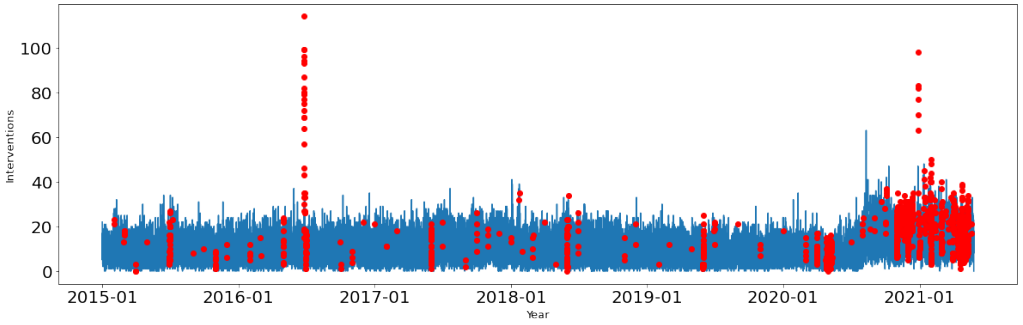


Figure 6.9: Anomalies detection

6.4/ INTERVENTIONS PREDICTION

Two experiments were performed on the dataset after applying all the above steps.

The first experiment aims to predict the number of firemen interventions for any period or date. This requires replacing the detected anomalies with the mean of the non-outlier values, as shown in the Figure 6.10, then the breakpoint (presented in green dots) was recalculated in Figure 6.11 to test if the changes influenced the selection of the coronavirus period as the point of change in the dataset, and finally the interventions for the period before and after the peak are predicted in Figures 6.12 and 6.13.

On the other hand, the second experiment predicts the values considered abnormal, as depicted in Figure 6.14 by replacing the normal data points with their mean. The idea behind this is to test the efficiency of our model in predicting an abnormal value that is certain to deviate from the normal trend, particularly during the coronavirus pandemic.

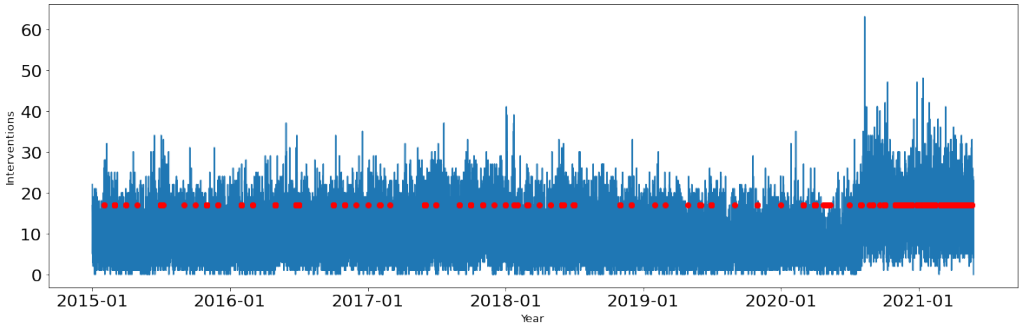


Figure 6.10: Dataset after replacing the anomalies

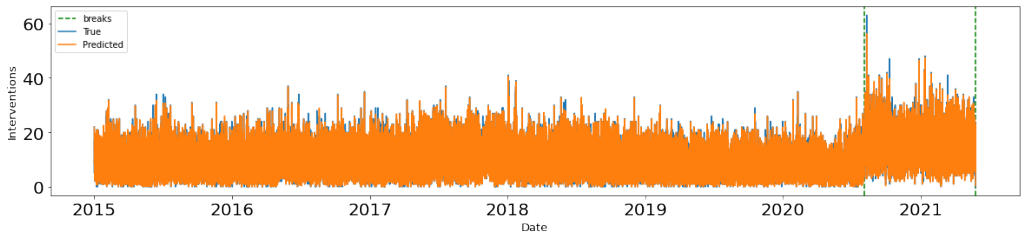


Figure 6.11: Breakpoint detection and interventions prediction after replacing the anomalies

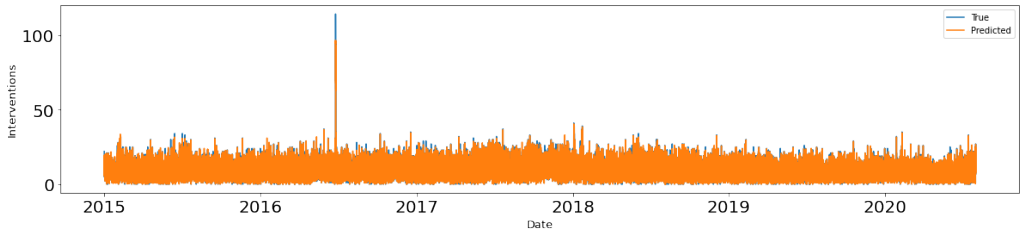


Figure 6.12: Period pre COVID-19 peak

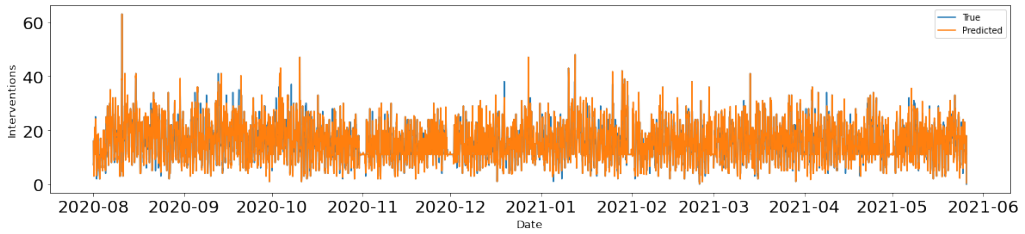


Figure 6.13: Period post COVID-19 peak

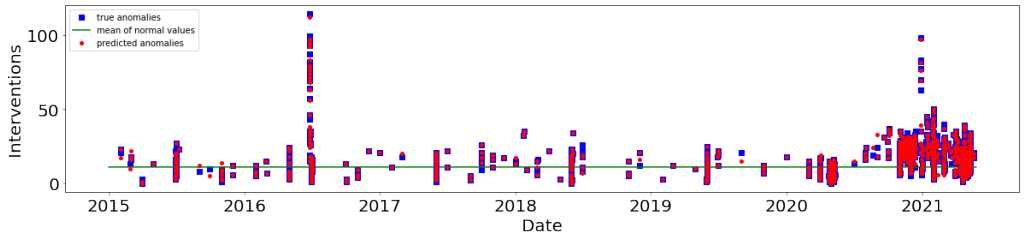


Figure 6.14: Real vs prediction for anomalies values

In addition, to check how accurate the prediction is for the test set, the period before August 5, 2020, and the period after August 5, 2020, the Mean Absolute Error and Root Mean Square Error metrics were used as indicated in Figure 6.15. Technically, these statistical characteristics show the error between the actual and predicted values.

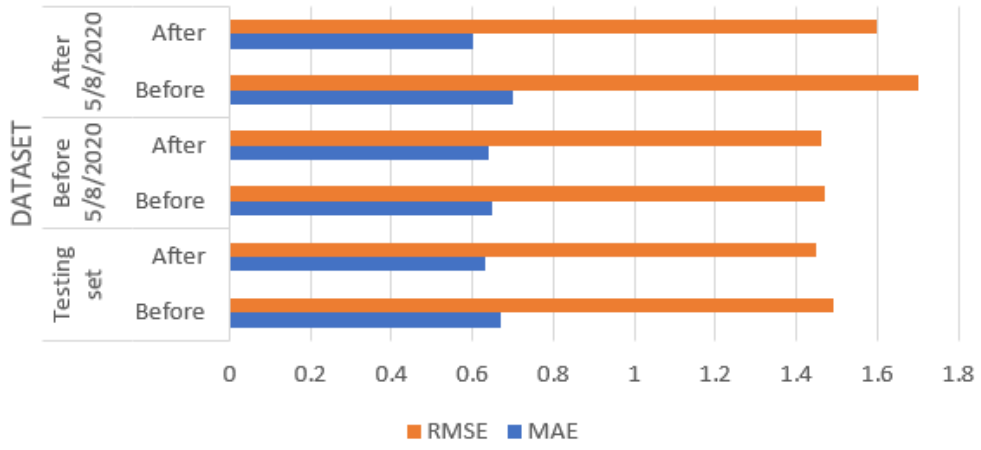


Figure 6.15: Statistical Features for the whole dataset, period pre and post COVID-19 peak

6.5/ DISCUSSION

Feature selection in this chapter was a turning point to overcome the challenge of a dataset with 1572 columns and to bring down the computational time by eliminating irrelevant attributes. This technique gets significant results as the root mean square error decreased from 2.192 to 1.49, and on the other hand, the training time was reduced from 230.382 seconds to 19.419 seconds.

Besides, it is highly suggestive that the breakpoint discovered on August 5, 2020 is directly related to the rise in COVID-19 cases in France, which has led to a significant increase in the number of firefighters' interventions. This peak was reached nearly three months after the initial lockdown, when the virus began to attack the younger generation. In this step of the analysis, the feature selection method was replicated for the periods before and after the breakpoint. As the results show, the importance scores of 'confinement2' and 'couvrefeux' were zero for the dataset before the peak of COVID-19 and increased thereafter. However, for 'confinement 1', the values were reversed as shown in Figures 6.16 and 6.17.

Here we see that the real pandemic crisis in France started as early as August 2020: the number of new cases reached an average of 3,003 per day, a figure four times higher than the average of 746 per day in July. This automatically leads to a higher demand for the sanitary service to accommodate all these patients, and thus to an increase in the number of firemen interventions.

A further step to improve model accuracy is to sort the firemen interventions into normal and abnormal values. After detecting 1122 anomalies and replacing them with the mean of the non-outlier values, better prediction was achieved for the entire dataset and for both

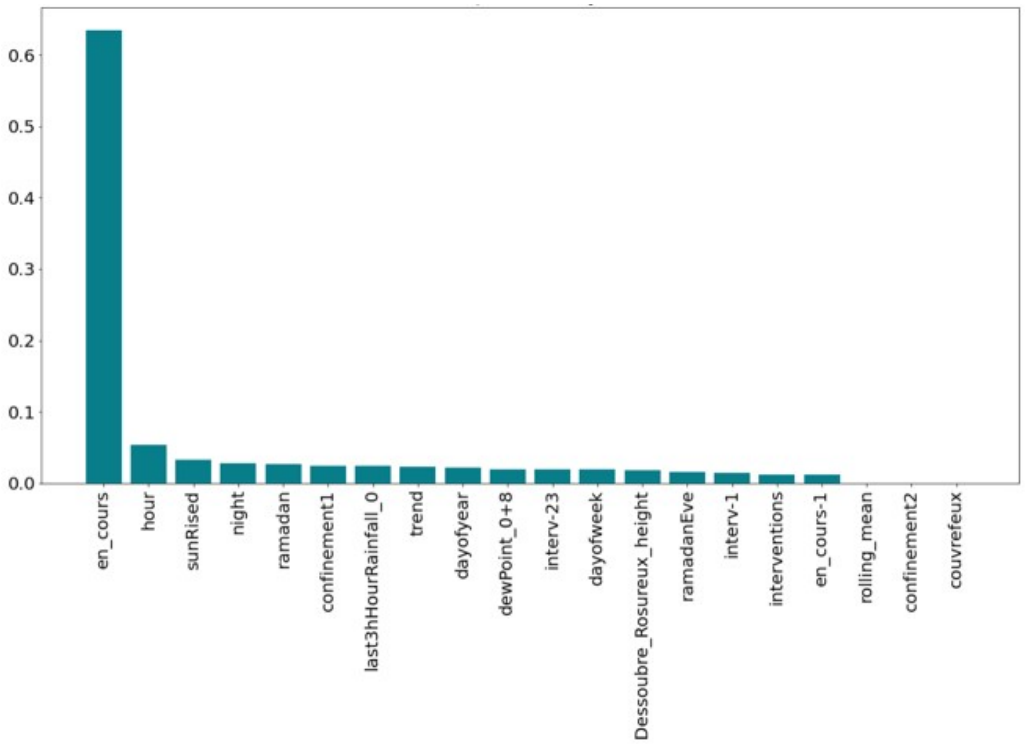


Figure 6.16: Feature importance before 5 August 2020

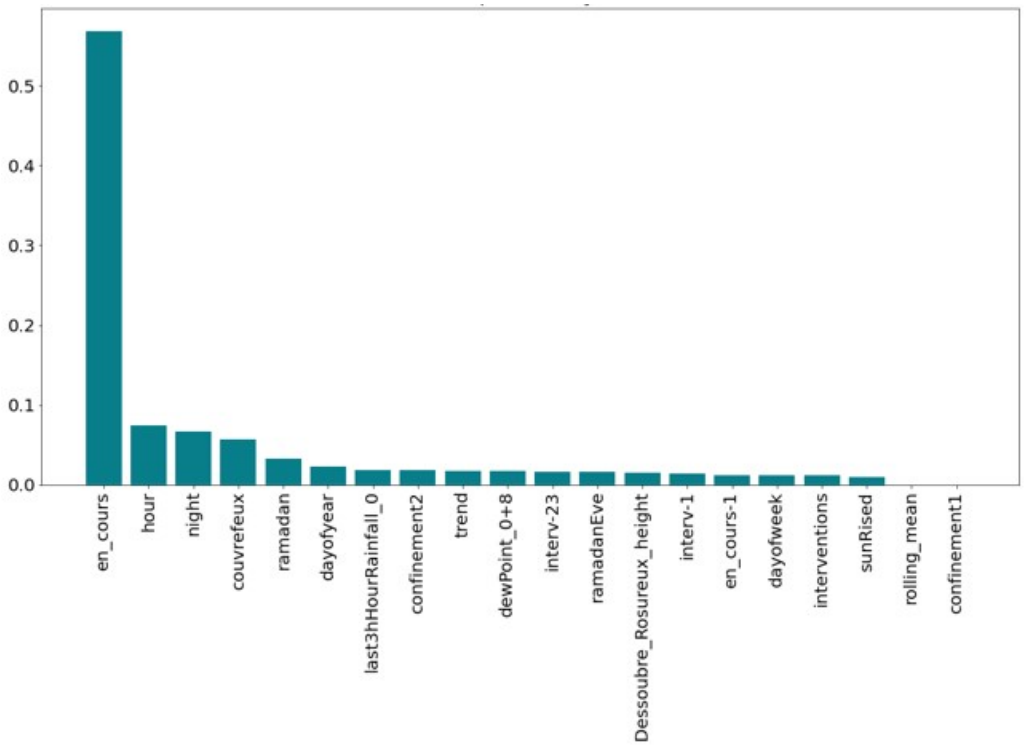


Figure 6.17: Feature importance after 5 August 2020

the pre and post COVID-19 periods (indicated by a reduction in MAE and RMSE).

On top of this, the most important result is that after applying modifications to the dataset,

the breakpoint always remains the same. This shows that it is impossible to ignore the coronavirus period. Even after replacing all anomalies, the COVID-19 period was always detected.

Nevertheless, it is a matter of great concern to verify the predictive accuracy of the interventions classified as abnormal. After replacing the anomalies with the mean of normal values, the results of MAE 0.762 and RMSE 1.995 are very prosperous.

6.6/ CONCLUSION

The aim of this chapter is to predict the number of times firefighters are called after detecting a breakpoint associated with the COVID-19 epidemic. The first step was to select features by applying the feature importance method. Then the breakpoint was detected and the last anomalies were replaced. In each step, the prediction of firemen interventions was performed using the XGboost algorithm. The results show promising accuracy in predicting regular or irregular events such as the COVID-19 epidemic.

7

TYPE OF FIREMEN INTERVENTIONS

This chapter focuses on predicting the target value of fire calls by creating 14 different subsets of data for each type of possible category. In previous chapters, the number of interventions was projected using different time periods, including the COVID-19 pandemic. What distinguishes this work from the previous ones is the completeness of the types of interventions in the dataset, rather than just the number of firemen needed. The methodology was based on the Departmental Fire and Rescue Doubs (SDIS 25) in France, which is identical to the one used in the previous chapters, but with a different time period. Adjusting the need to the demand of fire departments by categories of operations helps firefighters to be well prepared by knowing the type and number of human resources needed for the next operation.

7.1/ INTRODUCTION

Over time, firefighters' calls have increased, and now they are on strike in France, where they have been on the front lines fighting the coronavirus pandemic for the past three years. Therefore, with the use of ML is possible to predict the number of firefighters' missions, which can directly lead to relief and better control of financial, material, and human resources by estimating the possible flow of interventions in the next hour, the next day, the next week, the next month, and the next year. Consequently, such forecasting will improve the efficiency of emergency operations while reducing financial operating costs.

The purpose of this study is not only to predict the number of firefighters' missions over time but also to include the type of operation in the prediction (i.e., birth, fire, suicide, etc.). Hence, the number of available firefighters should be synchronized with the needs and requirements of each type of intervention. The dataset used in this study contains information on firemen operations registered by the fire and rescue department SDIS 25 by blocks of one hour in the Doubs-France region from "01/01/2015 00:00:00" to "31/12/2020 23:00:00".

This chapter describes how the dataset provided by SDIS 25 was sorted by 14 different types of interventions, resulting in 14 new subsets. In addition, statistical features in predicting the number of fire interventions were calculated for each type by using XGBoost and LightGBM algorithms.

To boot, many researchers have explored machine learning techniques for modeling and optimizing emergency services related to fire departments, some using basic statistical models and others using advanced machine learning as stated in Section 2.2.

Two different approaches were implemented in this chapter. In the first approach, 14 subdatasets were created for each intervention type, and in the second approach, both the type and number of intervention datasets were merged to gain more insights and to test the efficiency of such merging.

7.2/ CATEGORIES OF FIREFIGHTERS' INTERVENTIONS

The job of firefighters is not limited to knocking down fires and fighting forest fires. In all, 14 distinct categories of intervention are possible in each fire brigade mission.

1. Childbirth: delivery of a baby imminent or in progress in a public or private place.
2. Fire: any kind of fire in buildings, homes, businesses, industries, forests, trash, or any fire in the means of transportation (bus, boat, truck, train, tram), etc.
3. Fire on public road: the same concept as in the category 'fire', but on public roads.
4. Suicide: for any reason and attempt.
5. Traffic accident: by a transport vehicle or pedestrian.
6. Drown: in the swimming pool, or while researching someone died in the water.
7. Water-flood: any miscellaneous operations caused by floods.
8. Heating: any arson (single or group), any fire detected by smoke or fire alarm and any fire in an industrial or residential building, etc.
9. Emergency aid to people: any urgent mission, such as a cycling accident, paragliding, parachuting, delta plane, skiing, weapons, logging, hunting, recreation, spelunking, sports, and work. Also childbirth, interruption of cardiac and respiratory breathing, asphyxiation, burns, falls, etc.
10. Help for people: same concept as in the previous category, but not urgent, such as help for ambulance, help for person, pain, depression, trauma, search for missing person, etc.

11. Public road accident: any accident caused by a vehicle on the public highway.

12. Brawl: any fight between two or more persons with or without weapons.

13. Witness: in private or public, causing unconsciousness or involving difficulty in breathing.

14. Wasp: destruction of hymenoptera or any kind of insect.

7.3/ DATA EXPLORATION

The fire and rescue department in Doubs, France, has established 14 different categories for firefighters' operations. Each category includes several types of missions. The dataset contains 76685470 rows presenting records from "01/01/2015 00:00:00" to "12/08/2021 08:00:00". In this study, the selected data covers the period from the beginning of 2015 to the end of 2020, and the remaining records were omitted.

The dataset includes attributes related to:

- Date: formatted as "mm/dd/yy hh:mm:ss", which indicates the exact time of the operation
- Id: a unique number for each category of intervention
- Start and end: the start and end time of the firemen service operation
- Center: the location of the firemen department where the rescue was requested
- Reason: the category of the intervention
- in_progress: true or false
- Geom: the geometry where the intervention took place

It is logical and important to note that at the same time (hour, day, month, and year), several operations are possible: a woman might give birth to a child in one place, while a fire or an accident occurs in another.

Withal, the Fire and Rescue Department provided not only a record of the number of interventions but also another dataset that includes the type of interventions for each hour. In previous studies on the same topic, as stated in the Section 3.2 the datasets were conducted separately (by number or type of intervention) and never combined, as was the case in this chapter.

The dataset by intervention type draws on the same time period and step size as the dataset that carried out the number of interventions, but is much smaller. The number of attributes is limited to less context, including the reason or type of intervention. The comparison between the two datasets is shown in Table 7.1. For ease of naming, we refer to the first dataset as the “dataset number” and the second dataset as the “dataset type”.

It is also important to note that the “number dataset” has a unique line with the same index date, whereas the “type dataset” could have redundant indexes within the same date. This is because the first dataset indicates the number of interventions per day, while the second specifies the type of interventions per date, which may be different. For example, a possible scenario is to have 23 firemen operation on a given date and that these interventions are for different reasons and with different objectives, such as fire, accident, etc. So, as can be seen in table 7.1, there is a notable difference between the size of the two datasets in terms of attributes and rows.

	Dataset number	Dataset type
Duration	“01/01/2015 00:00:00” to “31/12/2020 23:00:00”	
Step period	1 hour	
Attributes size	1570	10
Rows size	52608	5886052

Table 7.1: Comparison between Dataset number and type

7.4/ PART1: CREATION OF 14 SUBDATASETS FOR EACH TYPE OF INTERVENTIONS

7.4.1/ SUB-DATASETS MODELLING

The goal of modelling new sub-datasets is to create a separate dataset for each type of intervention, rather than combine them all into one. First, the data is grouped by type of category (called reason in this study), resulting in 14 different subdivisions. Second, each sub-dataset is handled in a different file by transforming the existing dataset described in Section 7.3 into a new, meaningful dataset. The process begins by counting the number of deployments, grouped by type of mission, on the same date and time. This creates a new column called ’target’ which shows briefly the number of firefighters’ interventions by type and time.

Afterwards, columns related to the date are then created by accessing the values of the series using Pandas in Python on the Jupyter notebook and returning various properties,

such as:

- year
- month
- day number in the year
- days in a week
- hour in a day

An example of one dataset (Childbirth) is illustrated in Figure 7.1, after the above changes have been made. As can be seen, the first column refers to the date of the birth delivery interventions, with a target value indicating the number of firefighters who were involved in that call. The remaining columns show features related to the date of each intervention.

Figure 7.1: Childbirth sub-dataset

Date	target	year	month	dayOfYear	dayOfWeek	hour
1/1/2015 1:00	7	2015	1	1	3	1
1/1/2015 2:00	7	2015	1	1	3	2
1/1/2015 3:00	0	2015	1	1	3	3
1/1/2015 4:00	0	2015	1	1	3	4
1/1/2015 5:00	0	2015	1	1	3	5

On the other hand, Table 7.2 reveals the size of each resulting sub-dataset.

Also, Figure 7.2 shows the trend of the childbirth sub-dataset after completion. The same was done for the 13 remaining subsets of data, but not all plots were included in this paper.

7.4.2/ SUB-DATASETS APPRAISAL

After modeling and processing all sub-datasets, the prediction of the number of firefighters' interventions was performed using XGBoost and LightGBM. The hyperparameters for each algorithm were selected using the Optuna optimization system, and at each step, statistical features, particularly the mean absolute error and root mean square error, were calculated. Finally, after the 14 experiments were completed, a further investigation was conducted for each subset of data to assess the feasibility of using the type of interventions in predicting firefighters' operations as represented in Figure 7.3.

The details of each step are described as follows: an empty dataset named dfTotal was created, indexed by a date range similar to that used in this study, i.e., from the beginning

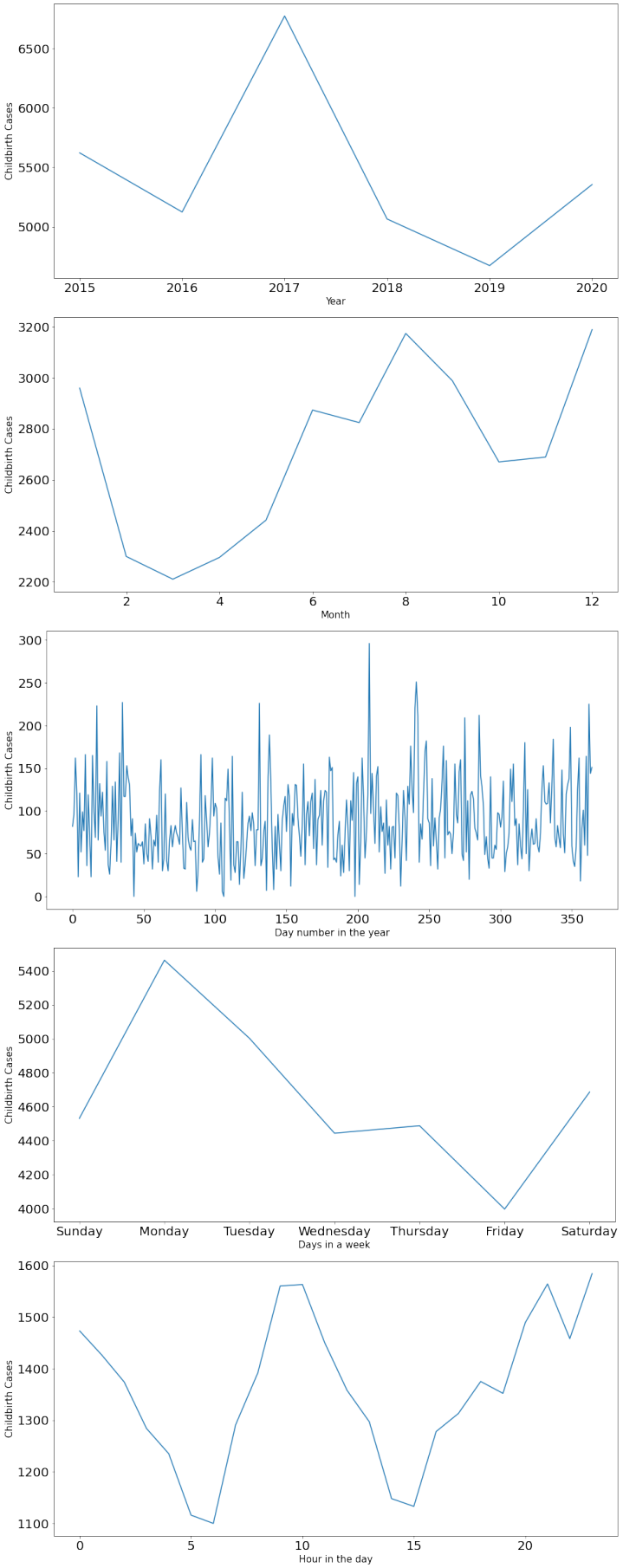


Figure 7.2: Time/date aspects of childbirth sub-dataset

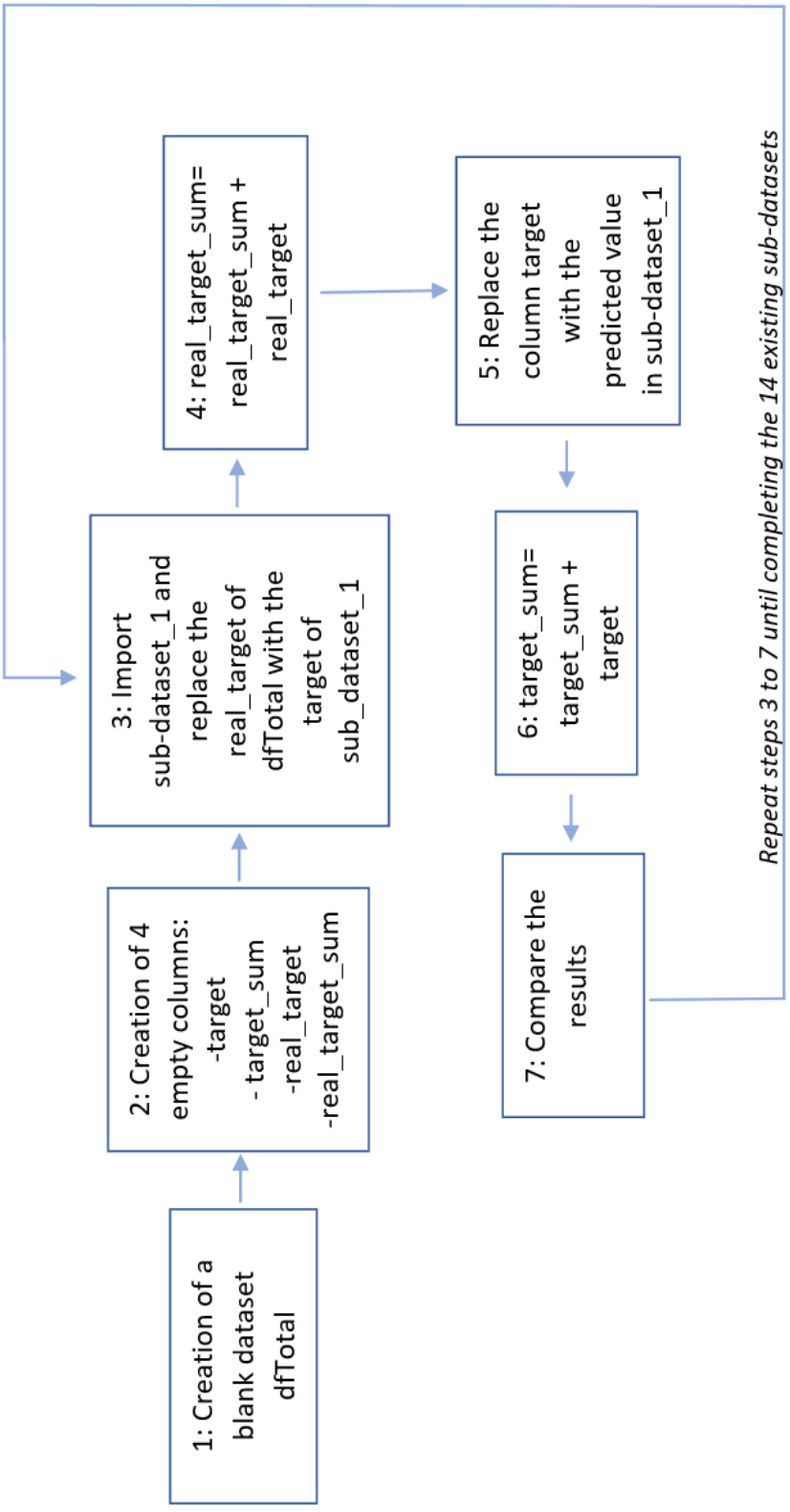


Figure 7.3: Steps of experiments

Category	Description	Dataset size
Childbirth	Labour/delivery	36905
Drown	Submerge/flood	39203
Wasp	Insect stings	55045
Brawl	Rough fight	66161
Fire on public road	In any public location	115604
Suicide	Dying intentionaly	157019
Water-flood	Water submerging	167880
Public road accident	Highway/train/bus...	732584
Traffic accident	Vehicles accidents	737529
Witness	Unconsciousness	738327
Heating	Arson, fire detected by smoke, etc...	924381
Fire	At home, bulduings, ...	1299319
Help for people	Any non-urgent mission	1556440
Emergency aid to people	Any urgent mission	2410564

Table 7.2: Datasets by category by size ascending order

of 2015 to the end of 2020 in 1-hour increments. Then, four columns were added to the dataset, filled with zero values:

- Target: presents the predicted value of firemen interventions for each sub-dataset independently
- Target_sum: sums the total number of predicted interventions for all partial datasets
- Real_target: represents the number of interventions for each sub-dataset independently
- Real_target_sum: accumulates the total number of interventions for all the sub-datasets

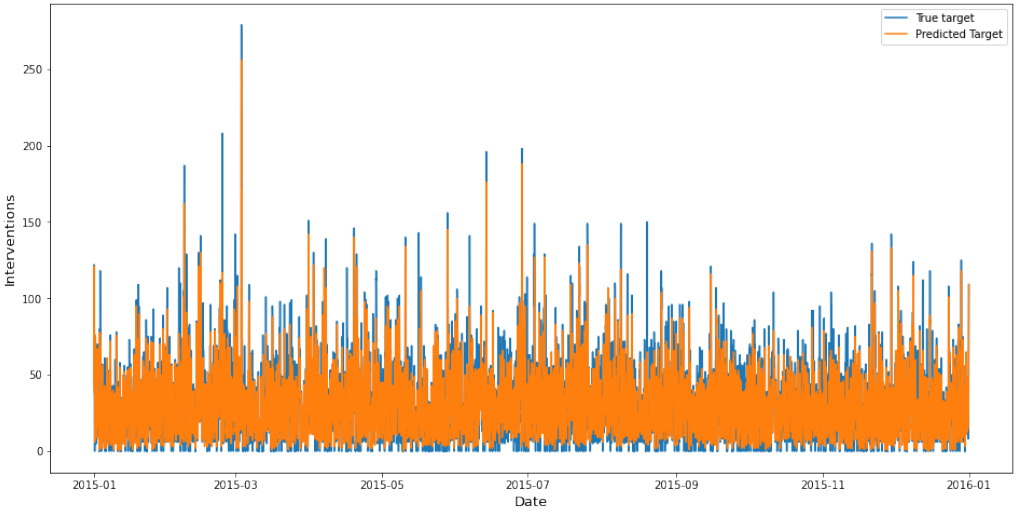
Both target and real_target are reset to zero after the result is retrieved for each sub-dataset The final result of dfTotal after execution of all sub-datasets is illustrated in Figure 7.4.

After training, validating, and testing each of the 14 datasets, the real versus the predicted number of firemen interventions is shown in Figure 7.5. However, MAE and RMSE for XGBoost and LightGBM for all the sub-datasets are represented in Figures 7.6 and 7.7.

Figure 7.4: Final result of dfTotal after execution of all sub-datasets

Date	target	target_sum	real_target	real_target_sum
1/1/2015 0:00	43	55	43	55
1/1/2015 1:00	108	158	119	179
1/1/2015 2:00	121	252	122	213
1/1/2015 3:00	75	272	75	276
1/1/2015 4:00	73	255	63	252
...	...	...	...	...
12/31/2020 19:00	61	312	61	313
12/31/2020 20:00	51	240	19	233
12/31/2020 21:00	26	198	26	198
12/31/2020 22:00	43	244	43	242
12/31/2020 23:00	30	245	30	245

Figure 7.5: Predicted vs Real target from 2015 until 2020



7.5/ PART 2: MERGING TYPE AND NUMBER DATASETS

7.5.1/ DATA RE-SAMPLING AND PROCESS

Working on the dataset by type of intervention, which has many categories on the same date, does not yield relevant information. Therefore, reassembling and reorganizing this dataset was the first modification to carried. We grouped the dataset by type of intervention (fire, delivery, etc.) and then created 14 sub-datasets for the different categories. In the next phase, the following steps were performed:

1. “Datasets type” for each category were trained and tested using LightGBM and XGBOOST. These subsets contain only information about the ’year’, the ’month’,

Figure 7.6: MAE using XGBoost and LightGBM

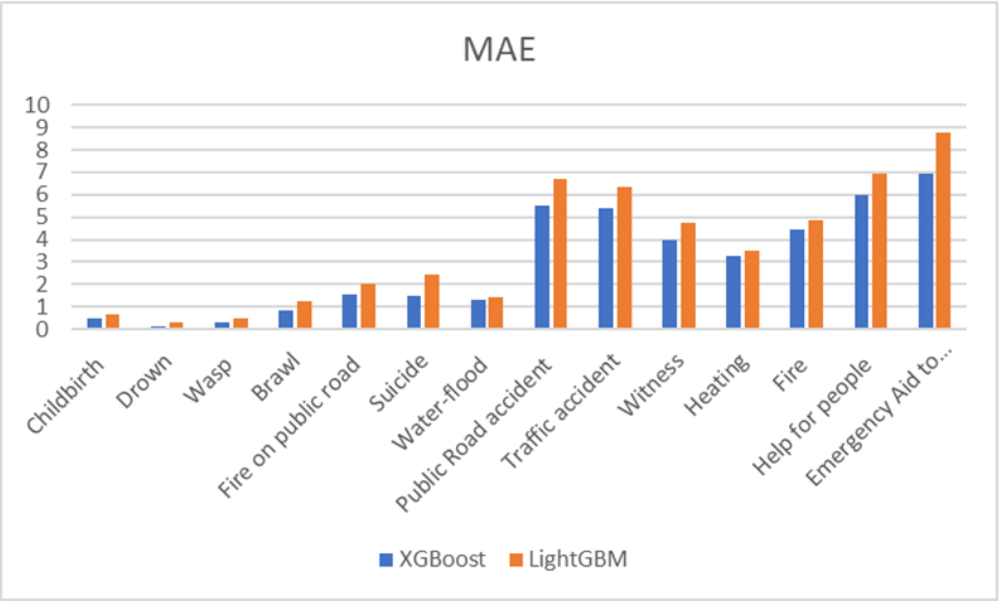
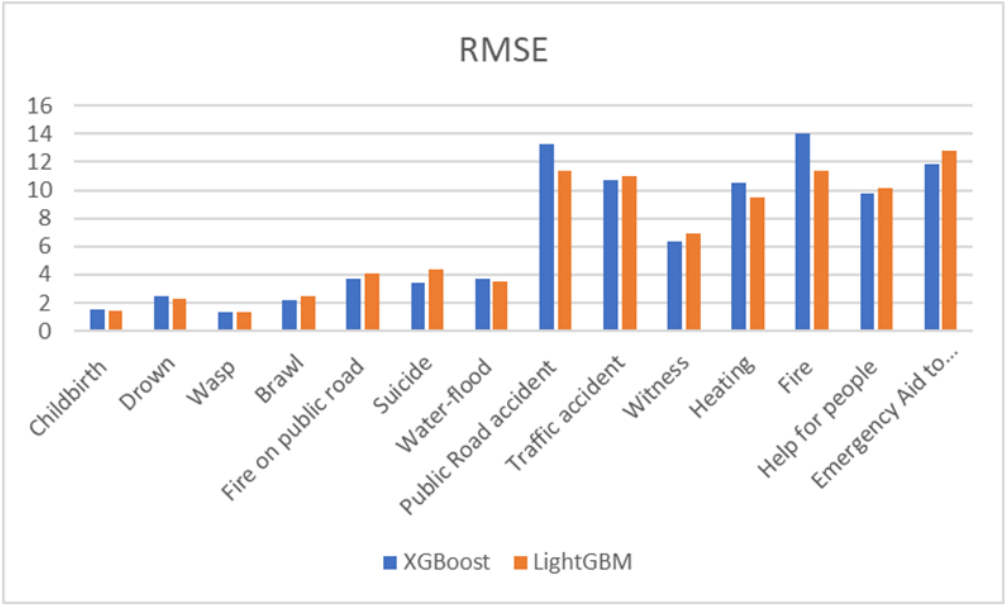


Figure 7.7: RMSE using XGBoost and LightGBM



the 'number of days in the year', the 'days in the week' and the 'hour in the day'.

2. "Datasets type" were merged with "Dataset number" to test the efficiency of adding more explanatory variables for weather, climate, vacations, occasions, etc.

3. Since step 2 requires a lot of computation time, feature selection was applied using XGBOOST featureimportance technique to reduce the computation time.

4. The accuracy was checked after selecting the best features of the merged dataset.

5. Results were compared all together to draw a conclusion.

7.5.1.1/ MERGING DATASETS

Similar to Part 1, XGBOOST LightGBM was used, and the values of each parameter were different for each category of firefighters' operations because the dataset was completely different. The combination of the two datasets, "number" and "type", was an important process for data analysis to predict the number of fire calls. The approach was to merge the "type dataset" and the "number dataset" into a new one using the index column carrying the date/time information. If the date was missing from the "type dataset" because there was no deployment associated with that particular date/time, a row with zero interventions was attached. Adding this row was essential in this study to allow a fair and accurate comparison across all days for all approaches.
To perform this merging process, we first ensure that both datasets have the same size, i.e., the same date range. Second, we omitted the 'target' column from the "number dataset", which refers to the number of deployments, since the purpose is to predict the number of fire deployments per category. Therefore, the 'target' considered in the coming experiments is the one included in the 'type dataset'. We also verify that the format of the indexes is identical in both datasets. Table 7.3, Table 7.4 and Table 7.5 show a selection of the original dataset 'Fire', the type of dataset, and the resulting dataset after the merging process, respectively.

Date	target	Year	Month	Day in the year	Day in the week	Hour
1/5/2015 5:00	7	2015	1	5	0	5
1/5/2015 6:00	7	2015	1	5	0	6
1/5/2015 7:00	0	2015	1	5	0	7
1/5/2015 8:00	5	2015	1	5	0	8

Table 7.3: Fire Dataset sample for 4 consecutive hours

7.5.1.2/ FEATURE SELECTION

The 14 merged datasets evidently contain a large number of attributes, with 1570 attributes coming from the original "number dataset" and 6 more coming from the dataset of any category of intervention, i.e., a total of 1576 attributes. These superimposed features require a lot of computation and training time. It took too many resources and was type oflly not convincing. Many attributes are irrelevant, and their presence in the dataset

Date	rolling mean	current Weather_0	...	noon	night	daylight SavingTime
1/5/2015 5:00	-1.04268	0.230316	...	FALSE	TRUE	TRUE
1/5/2015 6:00	-0.8858	0.230316	...	FALSE	TRUE	TRUE
1/5/2015 7:00	-0.10139	0.230316	...	FALSE	TRUE	TRUE
1/5/2015 8:00	-0.10139	0.230316	...	FALSE	TRUE	TRUE

Table 7.4: Sample of the “Number Dataset”

Date	rolling mean	current Weather_0	daylight SavingTime	...	Month	Day in the year	Target
1/5/2015 6:00	-0.8858	0.230316	TRUE	...	1	5	7
1/5/2015 7:00	-0.10139	0.230316	TRUE	...	1	5	0
1/5/2015 8:00	-0.10139	0.230316	TRUE	...	1	5	5
1/5/2015 9:00	-0.41515	0.230316	TRUE	...	1	5	5

Table 7.5: Sample of the dataset after merging the Fire with the “Number Dataset”

does not play a positive role in prediction. The opposite is also true: when the number of variables is significantly high, the accuracy of prediction decreases. Therefore, in this work, the selection of a reduced number of attributes is desirable for practical reasons. To achieve this goal, the technique of feature importance with gradient boosting was applied, in which an importance score is calculated for each attribute, allowing them to be ranked and compared. For each of the 14 available datasets, the minimum number of features giving the best Mean Absolute Error and Root Mean Squared Error was chosen.

7.6/ EXPERIMENTAL RESULTS AND INTERPRETATIONS

From the first approach, the use of machine learning to predict the number of firemen interventions by type of mission has been explored. Two algorithms were used after creating 14 sub-datasets from an original one provided by SDIS 25 in Doubs, France. Statistical features were also calculated for each subset of data to verify the realism of these experiments. As shown, the overall accuracy of this work is promising. The representation of the predicted target is reasonable compared to the real interventions. Moreover, the MAE and RMSE for each type of firemen missions show that the prediction accuracy depends on two criteria: the type of intervention and the size of the sub-dataset. As can be seen, the errors of the dataset for “childbirth” are vanishingly small compared to the errors of the “emergency aid to people”. It is very logical to relate the probability of error to the size of the dataset in this case.
However, this does not apply to other sub-datasets such as “help for people” and “fire” which are considered large datasets with more than 1300000 attributes. In this condition, seasonality played a major role. Specifying the number of fire missions is much more feasible than predicting the number of missions to help people, since it makes sense to associate fire outbreaks with hot weather. However, there is no clear correlation between date/time and requests for help.

On the other hand, XGBoost performs better than LightGBM in most cases when comparing MAE and RMSE for both. Also, it is worth noting that the error increases as the size of the dataset increases. This is very obvious since the explanatory variables in this study are very limited and only consider the year, month, number of days in a year, days in a week, and hours in a day.
It is also evident that redundancy is possible in this work. For example, a fire in a building can be divided into the categories of ’fire’, ’heating’ and ’emergency aid to people’. For this reason, comparison with previous work is not possible, because the total number of operations per hour is completely different from existing work.

On the flip side, in approach 2, after processing the data and then merging them into datasets for each category of fire operations, and after reducing the size of the large datum by selecting the minimum number of characteristics that yield the best accuracy for the statistical features MAE and RMSE, the next phase was to make predictions and test the accuracy and errors for different approaches. For all experiments, the dataset was divided into train, validation, and test, with an early stop round of 20 for XGBOOST with a “poisson” objective function and 500 for LightGBM with a “gbdt” boosting type.
Briefly, three major experiments were conducted in this study, and statistical characteristics were calculated for each.

- original datasets of the 14 different categories containing only the 5 attributes (year,

month, number of days in the year, days in the numbers, and hour in the day)

- 14 datasets after merging the “type dataset” with the “number dataset”, deriving a huge datum of 1576 attributes
- 14 datasets after selecting the best attributes that provide the highest accuracy

MAE and RMSE were calculated for both XGBoost and LightGBM (Table 7.6), with the exception that after merging both datasets without feature selection, the experiment was performed only with XGBOOST because of the enormous computation time (Table 7.7).

As might be expected, feature selection resulted in a reduction in MAE and RMSE for most categories of the datasets. This explains that this technique is feasible and gives good prediction results, as shown in Table 7.6. Nevertheless, this need not be the case when forecasting a simple dataset containing only time and date attributes, as shown in Table 7.7. Dealing with complex and large datasets increases variability and thus the error rate.
Furthermore, comparing the statistical features of XGBOOSt and LIGHTGBM, it is conspicuous to state that LightGBM has the lowest MAE and RMSE. It is obvious that LightGBM is faster than XGboost, especially when the data is very large, as in this study.

	Before merging the datasets				After Feature Selection			
Dataset	XGBOOST		LightGBM		XGBOOST		LightGBM	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Childbirth	0.479	1.523	0.632	1.452	0.434	1.611	0.403	1.263
Drown	0.118	2.453	0.327	2.327	0.223	3.244	0.206	2.68
Wasp	0.322	1.313	0.471	1.332	0.417	1.851	0.403	1.274
Brawl	0.865	2.221	1.263	2.437	1.398	2.915	1.452	2.725
Fire on public road	1.577	3.688	2.049	4.042	1.509	4.911	0.926	3.186
Suicide	1.491	3.464	2.465	4.367	2.302	4.708	1.096	3.469
Water-flood	1.331	3.731	1.445	3.487	1.589	4.487	1.085	3.235
Public road accident	5.503	13.307	6.721	11.37	6.382	11.719	3.505	8.893
Traffic accident	5.406	10.689	6.348	11.029	6.567	11.806	3.332	9.423
Witness	3.967	6.402	4.775	6.932	4.292	6.76	2.258	5.64
Heating	3.247	10.546	3.507	9.485	4.14	16.908	4.768	17.234
Fire	4.435	14.07	4.848	11.374	4.93	22.406	5.036	28.863
Help for people	5.973	9.812	6.934	10.177	5.792	9.828	3.225	8.144
Emergency aid to people	6.914	11.85	8.772	12.846	8.087	13.43	4.227	10.864

Table 7.6: MAE and RMSE for “Type datasets”

Over and above that, the nature of the dataset played a major role in predictive accuracy.

Dataset	Without feature selection	
	MAE	RMSE
Childbirth	0.689	1.795
Drown	0.302	2.89
Wasp	0.653	1.835
Brawl	1.375	2.669
Fire on public road	2.244	4.376
Suicide	2.507	5.203
Water-flood	2.061	4.35
Public road accident	9.935	14.958
Traffic accident	9.869	14.874
Witness	7.506	11.852
Heating	7.421	23.19
Fire	10.305	23.054
Help for people	11.256	16.545
Emergency aid to people	12.145	16.669

Table 7.7: MAE and RMSE after combining the datasets and before feature selection

In fact, the 14 datasets vary greatly in size, ranging from 36905 to 2410564, which is about 65 times larger than the dataset for childbirth. Furthermore, not only did size play a large role in the error rate, but also the category of the dataset. Floods, fires, accidents, and many other fire incidents are somehow related to time and date, i.e., seasonality. However, some other types of dataset, e.g., child births, wasps, etc., are not so easy to predict because the frequency of occurrence is not known as a function of season or time.

7.7/ CONCLUSION

In this chapter, two approaches have been considered: first, the prediction of the number of firemen interventions was conducted considering 14 different types of missions that a firefighter may be called to. All the studies proposed in recent years consider different metrics but never the type of call-outs. From the analysis of the results in this approach, it appears that the integration of the category of operation is feasible and provides accurate results. The subject of interventions and the size of the sub-datasets also played a major role in the accuracy. Secondly, an improvement of previous studies was achieved in approach 2 by merging the two datasets on the number of firemen interventions and

the type of missions provided by the Department of Fire and Rescue SDIS 25 in Doubs, France, in order to investigate the effect of adding explanatory variables to the existing attributes, which are considered very simple and refer only to the date and time. The datasets created in approach 1 were merged with the huge original data used in previous research, and finally, the best features for each sub-dataset were selected using the feature importance technique to reduce computation time and storage requirements. All the assessments were performed using two well-known boosting machine learning algorithms, XGBoost and LightGBM.

The results demonstrated good accuracy when processing the 14 sub-datasets of different categories without adding explanatory variables or feature selection. On the other hand, comparing the accuracy before and after feature selection, the errors were lower after the number of attributes was reduced, but never better than the original 14 sub-datasets before any change.

8

K-MEAN CLUSTERING FOR FIREMEN INTERVENTIONS

In this chapter, a new dataset was implemented, completely different from all other contributions, by considering the Île-de-France region instead of Doubs, France, which was limited to a small number of villages. The K-Means clustering method was applied, and further research was conducted to summarize the criteria by which the Insee were classified into clusters.

8.1/ INTRODUCTION

Cluster analysis is widely used in various fields to classify data that has similarities to objects in the same group but are different from the objects in the other group, to gain insights into various applications such as marketing, urban planning, fraud detection, biology, and many more. The ability to cluster the number of fire operations in the Île-de-France, especially in Yvelines, will definitely help the fire and rescue services to make better decisions in emergency response and increase the efficiency of material and human resources, which, if they can be reduced, can lower financial costs.

Dozens of studies have been conducted on this topic to analyze the same dataset and predict the number of fire calls. Therefore, the main objective of this study is different from the previous ones and is based on clustering the dataset to draw a useful conclusion that can improve the emergency response of firefighters. The ability to determine the need and demand in this sector has proven to be reliable when using different techniques. However, in this work, sorting and grouping the number of fire brigades into different clusters leads to a better understanding and interpretation of the fire departments in the region of Île-de-France.

8.2/ Related Work

Cluster analysis, which aims to discover groups in data, is used in numerous applications, including marketing, biology, geology, libraries, urban planning, document analysis, and many more. In business, clustering can help with customer segment discovery, especially since effective marketing today focuses on the customer, not just the product. Retailers need to target a set of customer segments that clearly express business value. A study conducted in [Raiter, 2021] considered cardholder data from different banks based on purchase frequency and income to maximize bank efficiency, service quality, and customer satisfaction. Another method, developed by [Ahani et al., 2019], collected data from travelers' reviews of wellness hotels on TripAdvisor to predict travel choices and segment spa-hotels to better develop spending on marketing strategies.

Additionally, in their work, [Tamba et al., 2019] aim to classify the books in Universitas Prima Indonesia library into different clusters: frequently, often, or rarely used/borrowed books. Their goal is to remove unused books to make room for others in the library and to bring more interesting books to readers.
In the field of urban planning, [Ip et al., 2010] in their study planned the installation of charging stations for battery electric vehicles in urban areas characterized by various complex factors such as traffic, small spaces, distribution of power grids, etc. by applying hierarchical clustering analysis.
Besides, document clustering is useful for search engine grouping, building document taxonomies, automatic categorization, and more. In a paper proposed by [Kuhn et al., 2007], linguistic information from source code such as comments and identifier names is retrieved by clustering the source artifacts with similar vocabulary.

On the top of that, many studies have shown the effectiveness of using machine learning in emergency response, especially the study dealing with fire department operations represented in Chapter 3, specifically in Section 3.2 and in all previous contributions. To the author's knowledge, there is no previously conducted study that summarises operations in the Île-de-France region. Literally all existing research on the same topic to date has been conducted on the dataset of the Doubs region, France.

8.3/ Materials and methods

8.3.1/ Repository overview

The Department of Fire and Rescue has expanded the data provided to include the Île-de-France region, whereas previously all data referred to the Doubs-France. It is important to note that the collection of such data by the SDIS department is not a simple process,

as it requires many calculations and storage.
In this study, the dataset contains information on fire calls from 1/1/2017 0:00 AM to 9/9/2020 9:55 AM with different attributes about the location of the call center, the start/end/alarm time of the call, the type of mission requested, and insee (the acronym for National Institute for Statistics and Economic Studies), which has numerical indexing codes for various entities in France. Insee also produces official statistics.

With all these attributes, neither time series forecasting nor clustering can be performed. The dataset must be cleansed of irrelevant features that are useless for such a technique. Therefore, the selected attributes are the start date, which is incidentally considered as the index of the dataset, the insee, and the type of intervention. Other attributes could be an important feature in future studies.
Besides, it is not the aim of this study to investigate the nature of the interventions. However, the reason why the attribute “type” was chosen is that the category of interventions was grouped by the insee and then the number of interventions recorded was counted. Therefore, the resulting dataset contains the date (index), the insee, and the number of deployments for each site on each date.
It is crucial to note that the index can be duplicated, as different interventions can occur for different insee. For example, on January 1, 2017, there were 3 interventions in insee 92022, but only one intervention in insee 78003. In addition, the dataset contains 307 different insee belonging to 6 different departments in Île-de-France.

Table 8.1 shows the total number of insee and interventions grouped by department. As can be observed, there is a large discrepancy between the numbers of interventions in the different departments, so performing clustering experiments for the entire dataset does not yield a convincing result. Therefore, the only department selected was ’Yvelines’, which has the highest number of insee and firemen missions. Moreover, only the 20 largest insee of Yvelines illustrated in Figure 8.1 were selected for the experiments in order to achieve a feasible clustering. A trimmed portion of the resulting dataset, used in the remainder of this article, is shown in Table 8.2.

Table 8.1: Total number of insee and interventions for each department

Department	Total Insee	Total interventions
Yvelines	259	495938
Hauts-de-Seine	10	2315
Eure-et-Loir	11	808
Essonne	7	222
Val-d’Oise	13	213
Eure	7	183

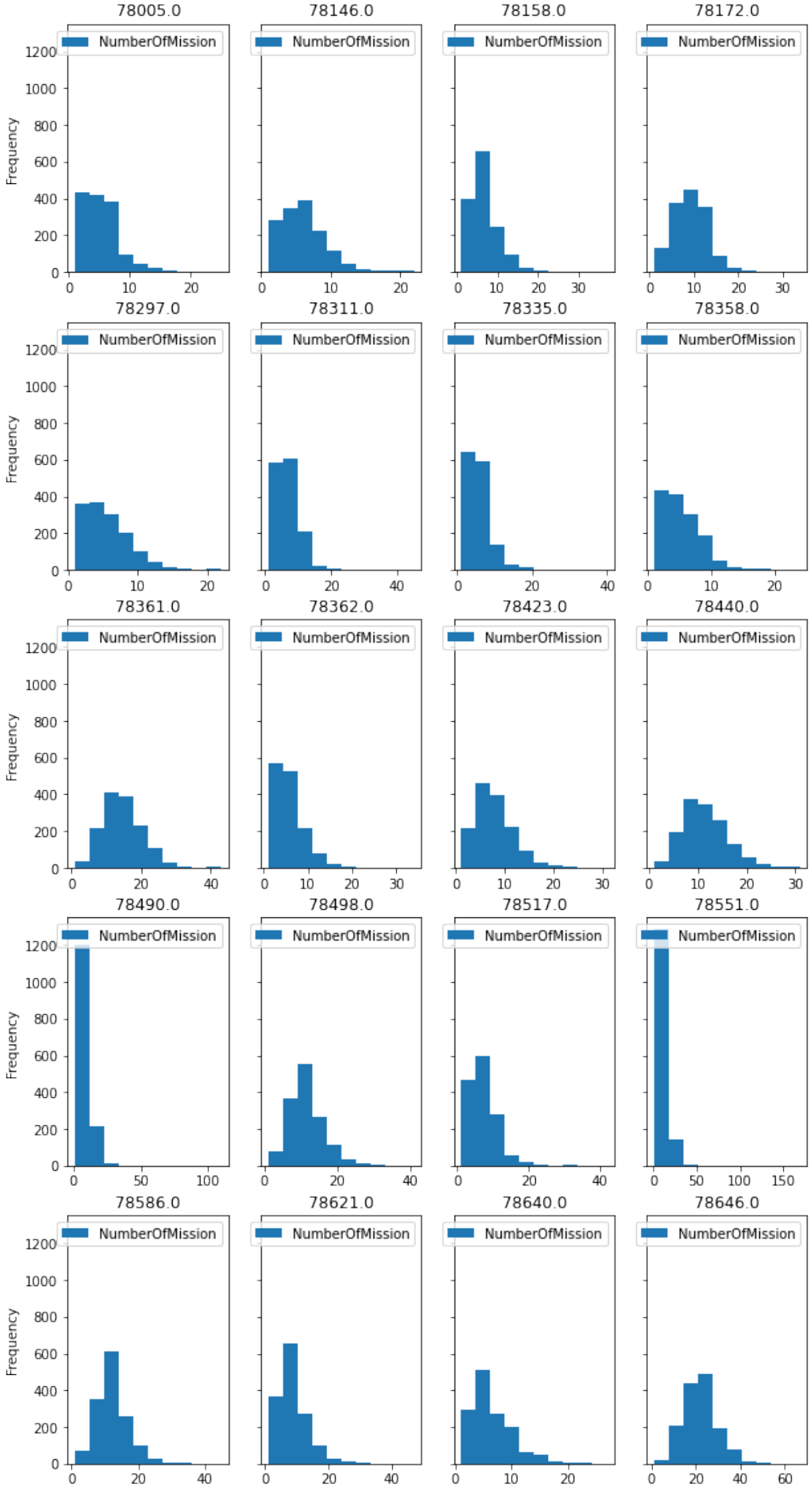


Figure 8.1: Frequency of each number of missions for all insee of Yvelines

Table 8.2: Yvelines Dataset

Date	**I**nsee	Number of Missions
2017-01-01	78005.0	3
2017-01-01	78146.0	7
2017-01-01	78158.0	10
2017-01-01	78172.0	11
2017-01-01	78297.0	7

8.3.2/ DICTIONARIES

After cleaning and preparing the repository, a data structure called a “dictionary” was first used to store the data in keys. It consists of key values known as an associative array that can be accessed by keys. Twenty dictionaries were created, and the corresponding number of fire brigades missions were assigned to them for each date. Then, the clustering procedure is developed using the different keys of the available dictionaries. As can be seen in Figure 8.2, the number of items that make up each dictionary is almost in the same range, which makes the clustering procedure more practical than keeping all departments and all insee in the experimentation.

8.3.3/ CLUSTERING TECHNIQUE

Clustering is a multivariate data mining analysis method that uses distance measurements to divide objects into homogeneous, disjointed classes called clusters based on object similarity. The variance of a cluster must be minimal, reflecting high similarity between the elements to which it belongs. In other words, elements with high similarity belong to one cluster, while elements with low similarity share different clusters. In this study, the Time Series Clustering technique uses only information about the date and the number of deployments, regardless of other attributes. The analysis was performed using the K-means clustering technique [Lloyd, 1982, MacQueen et al., 1967].

In particular, the main goal of clustering is to divide the dataset of firemen operations into groups that share certain similarities or common characteristics. The classification according to certain criteria can help the fire and rescue service to better identify and interpret the reasons and frequency of the calls, which will definitely help to meet the needs and increase the efficiency of the emergency response. The 20 created dictionaries, each referring to one insee, are shown in Figure 8.3, displaying some similarity between the diagrams. Moreover, some outliers can be seen, represented by a spike in the graph, which has been replaced by the average of the deployments of the firefighters

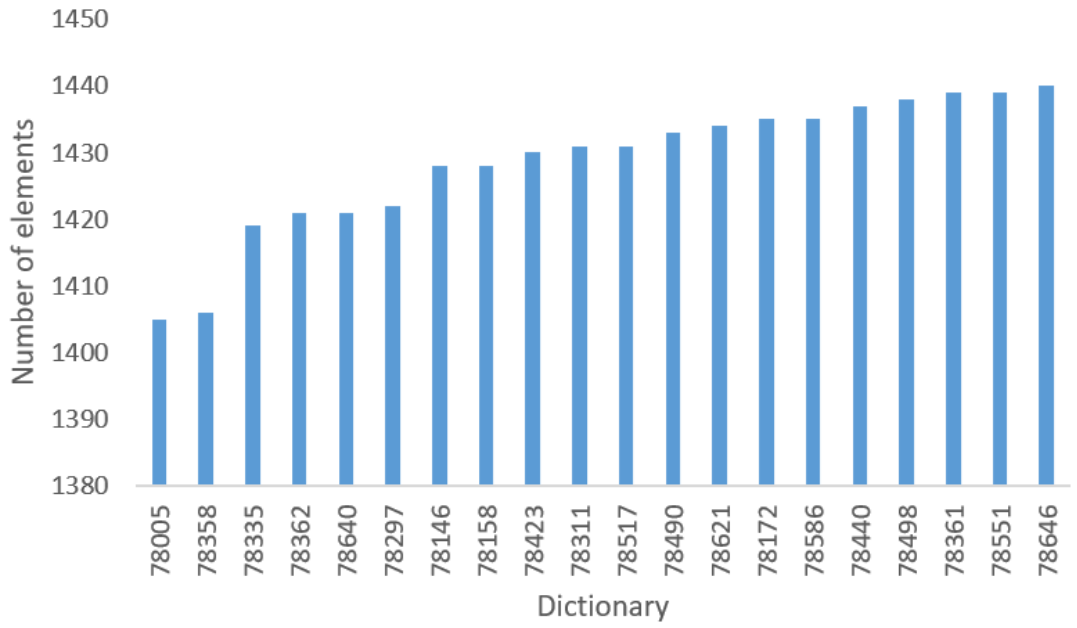


Figure 8.2: Number of elements of each dictionary

in the corresponding dictionary. As a consequence, both insee 78551 and 78490 have been changed, resulting in a more consistent graph.

The clustering method used in the experiments is the k-mean technique, in which the probability of the most pertinent function is calculated and the data set is divided into k clusters, keeping the clusters as separated from each other as possible and as compact as possible [Kaur et al., 2013, Bansal et al., 2017]. In K-Mean, centroids are first randomly determined for the clusters, and then the data points with the greatest similarity are assigned to the closest centroid. This loop is repeated until either the clusters no longer have a change or a certain number of iterations are completed.

Researchers have proposed many methods to find the optimal number of clusters for the k-mean algorithm (e.g., Elbow method [Marutho et al., 2018], Gap Statistic [Tibshirani et al., 2001], Cross Validation [Fu et al., 2020], Silhouette method [Lengyel et al., 2019] and many others). In this work, the optimal number of clusters was calculated by applying the elbow technique using “KElbowVisualizer”. The elbow technique runs the K-means clustering algorithm for a set of clusters (0 to 10 in this work), calculates the average score for each value of k for all clusters, and plots the variation.

The Figure 8.4 illustrates the elbow technique and shows the distortion score, which calculates the sum of the squared distance between each data point and the assigned centroid. To interpret the chosen k-value, another metric was used, namely the silhouette score. It measures the gaps between each data sample of the same cluster. It ranges from -1 to 1, and the closer the value is to 1, the better the clusters are separated from each other and the denser they are. A value of 0 reflects overlapping clusters. As can

be seen from the diagram in Figure 8.4, the optimal number of clusters for the k-mean algorithm is 3, with a dashed vertical line marking the “elbow”. Yet, after n = 3, the yield decreases as the k value increases and the line begins to become linear. On the other hand, the silhouette score found is 0.779, close to 1, meaning that the clustering quality is realistic.

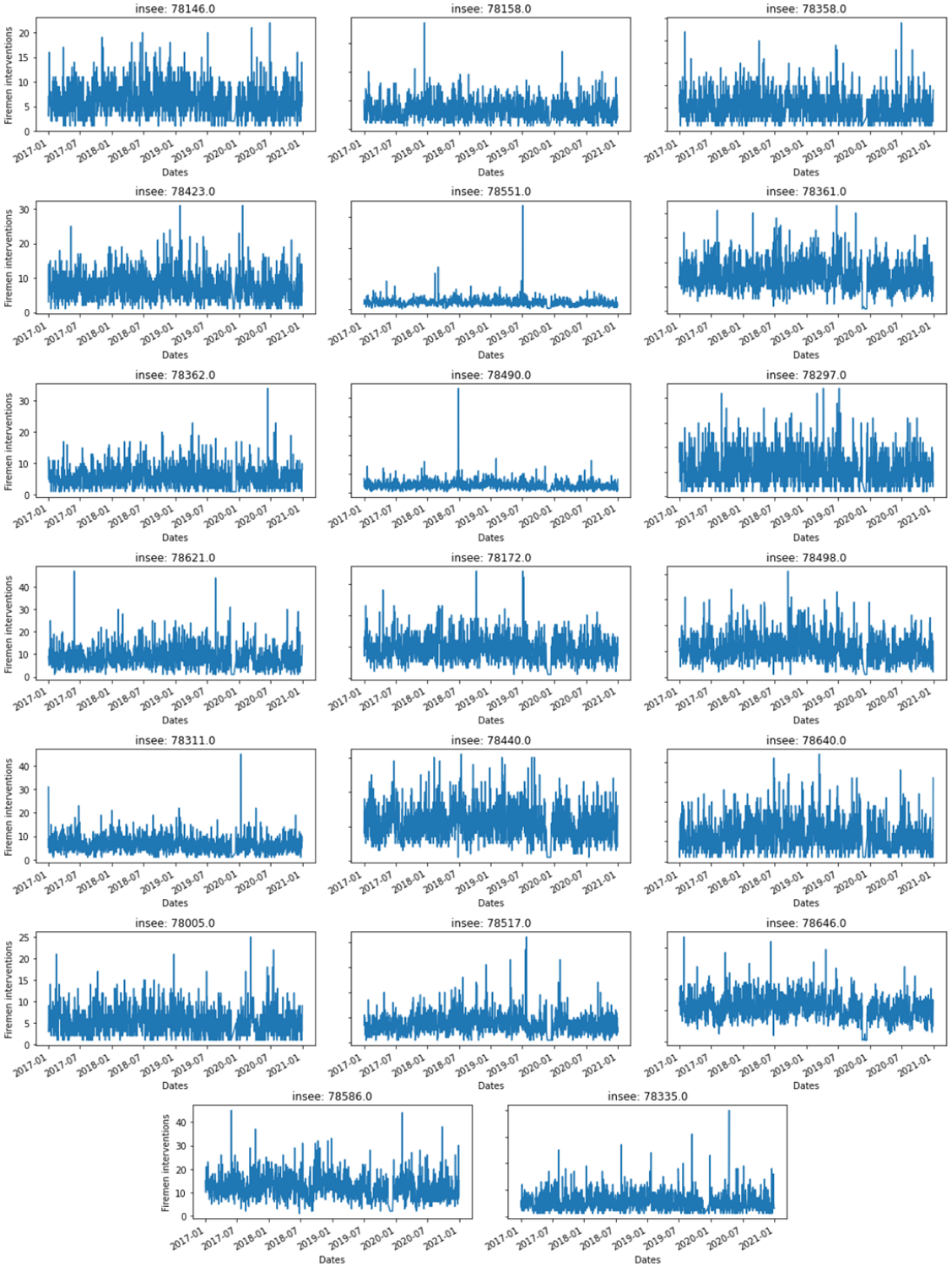


Figure 8.3: Number of elements of each dictionary

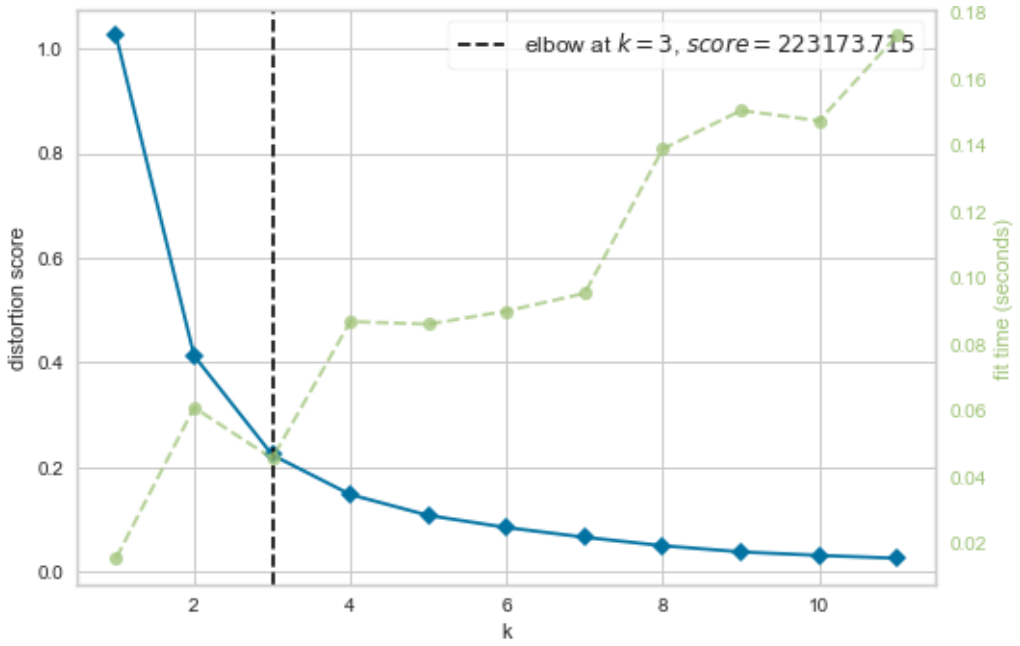


Figure 8.4: Distortion score Elbow for kmeans clustering

Along with the clustering technique and the optimal number of clusters, Dynamic Time Warping (DTW) was chosen as the distance metric for Time Series [Sakoe et al., 1978].

8.3.4/ CLUSTERING RESULTS

The twenty insee picked up in the department of Yvelines were grouped into three distinct clusters without overlap: Cluster 0 (Figure 8.5), Cluster 1 (Figure 8.6), Cluster 2 (Figure 8.7). Additionally, Linear regression [Yan et al., 2009] was implemented to elucidate the mean absolute error and the mean squared error for each different cluster displayed in Table 8.3.

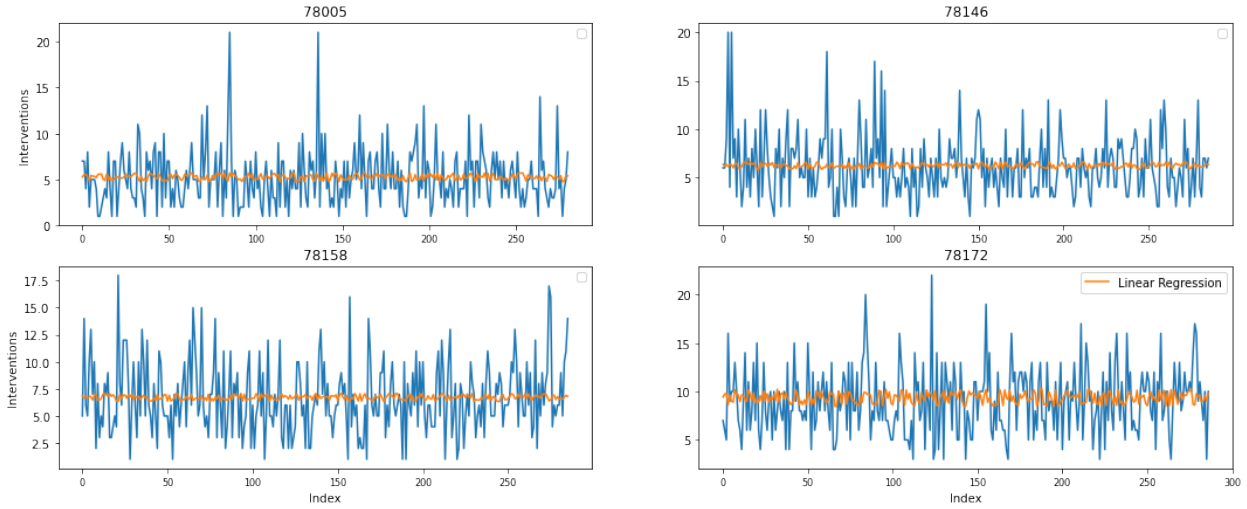


Figure 8.5: Cluster 0 composition

Figure 8.6: Cluster 1 composition

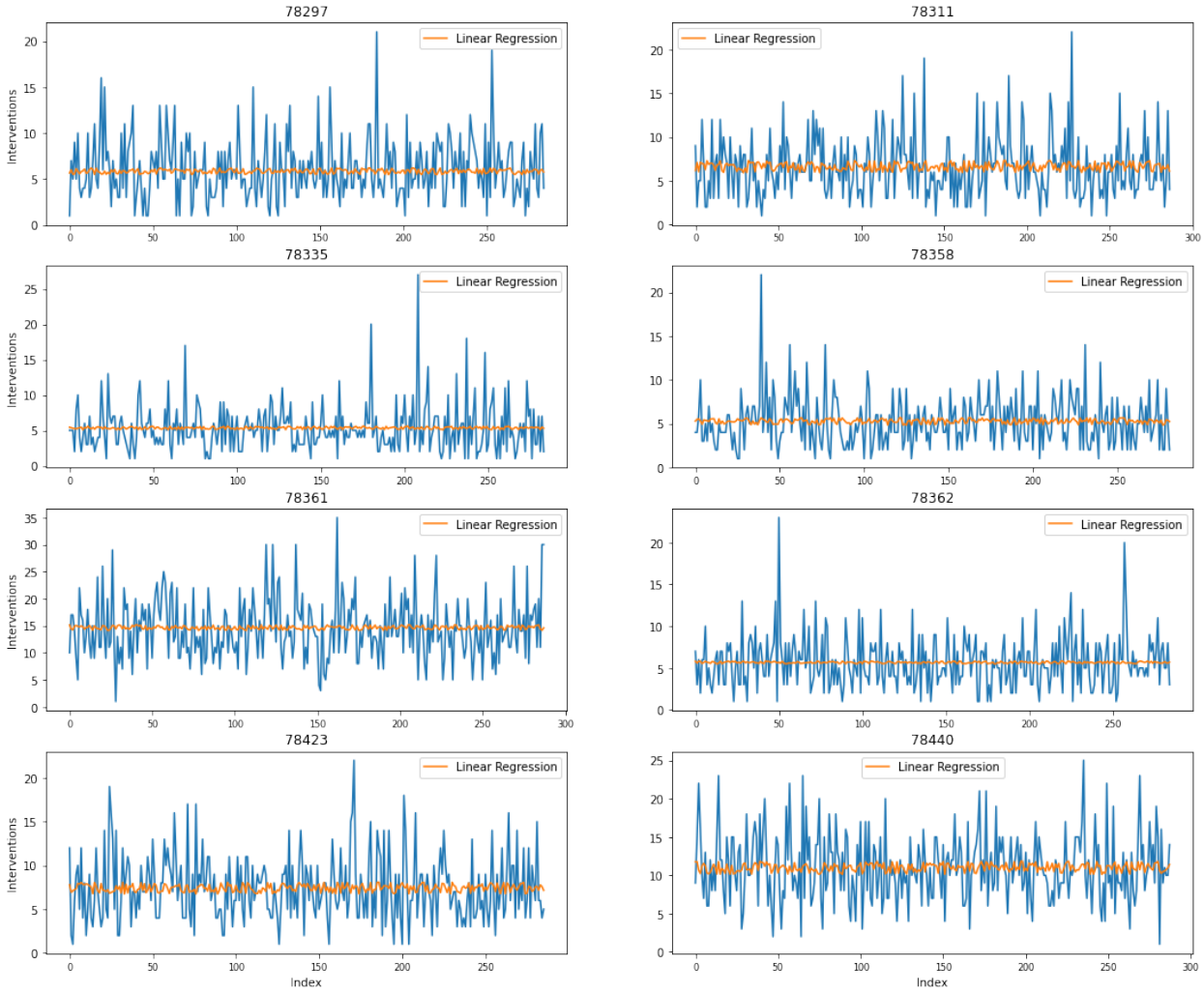


Figure 8.7: Cluster 2 composition

Table 8.3: MAE and RMSE for Linear Regression for all the insee

Cluster	insee	RMSE	MAE
Cluster 0	78005	3.050	2.343
	78146	3.197	2.455
	78158	3.323	2.654
	78172	3.412	2.732
Cluster 1	78490	3.983	3.050
	78498	4.545	3.367
	78517	4.447	3.240
	78551	5.380	3.891
	78586	4.728	3.729
	78621	4.729	3.366
	78640	3.707	2.893
	78646	7.486	5.797
Cluster 2	78297	3.329	2.629
	78311	3.444	2.650
	78335	3.345	2.429
	78358	2.820	2.168
	78361	5.579	4.405
	78362	3.119	2.419
	78423	3.725	2.968
	78440	4.380	3.514

8.3.5/ Further Investigations

Since this work only deals with time series that don't take into account explanatory variables other than the date, insee and number of firefighters' interventions, an analysis of the cluster distribution is necessary to verify which criteria were taken into account in the grouping process. Accordingly, after creating three different clusters for the 20 insee of the Yvelines department, the breakpoint was calculated using a library called 'rupture' [Truong et al., 2020] as indicated in Table 8.4. The reason for this step is to investigate whether the breakpoint affects the segregation of the clusters or not. More specifically, if a common breakpoint of clusters or insee is found, it justifies the metric of cluster partitioning.

Another study was considered, the statistics on the citizens of Yvelines: population size, people over 75 years old, and occupational groups (farmers, entrepreneurs, higher intellectual professions, employees), as well as pensioners and people with no professional activity (Table 8.5). This was evaluated to see if these statistics influenced the cluster breakdown.

Table 8.4: Two breakpoints detection for each insee

Cluster	Insee	Breakpoint 1	Breakpoint 2
Cluster 0	78005	3/11/2017	3/25/2019
	78146	9/12/2017	3/26/2019
	78158	8/29/2017	1/1/2018
	78172	8/28/2017	12/31/2017
Cluster 1	78490	11/27/2018	1/26/2019
	78498	12/1/2017	7/10/2019
	78517	9/14/2017	9/14/2017
	78551	10/7/2017	2/20/2019
	78586	9/22/2017	2/5/2019
	78621	2/11/2018	7/21/2019
	78640	9/3/2017	2/1/2020
	78646	7/10/2019	9/3/2019
Cluster 2	78297	8/25/2017	4/16/2019
	78311	10/8/2017	3/27/2018
	78335	7/8/2019	8/2/2019
	78358	7/6/2019	12/25/2020
	78361	6/25/2019	7/25/2019
	78362	10/24/2017	4/23/2019
	78423	9/16/2018	2/13/2019
	78440	9/3/2019	9/13/2019

Table 8.5: Statistics for distinct insee

Cluster	Insee 78-	Population	75 or older	Farmers	Entrepreneurs	Higher intellectual professions	Employees	Retired	No professional activity
0	005	21098	1051	3	292	1833	3511	2372	2249
	146	30330	2749	14	778	5965	3293	5078	3662
	158	31306	3914	3	619	3725	891	2978	1346
	172	35656	2998	9	664	3848	4743	6276	4322
1	490	31013	2209	3	522	3349	5274	3557	4232
	498	38313	3158	19	527	4475	5604	6951	4761
	517	26933	2703	22	347	3290	3487	5288	3232
	551	44750	4134	27	932	8660	5012	7590	7031
	586	52269	3800	8	1031	4706	8517	8699	6338
	621	32120	1180	4	312	910	5347	2 827	4471
	640	22649	1895	0	337	2856	3168	4287	1945
	646	85205	8811	31	1521	16756	10412	15211	14245
2	297	29332	864	0	302	4345	4676	1983	3425
	311	32449	2222	19	557	4943	4124	5013	3405
	335	17147	939	2	232	672	2435	2138	2538
	358	23611	2539	19	556	5187	1781	4578	2759
	361	44227	3098	0	531	1934	5932	6250	8879
	362	20499	1429	3	245	955	2754	3325	2648
	423	32575	1047	0	300	4454	905	2377	1775
	440	32949	1803	1	380	1108	5077	4575	6129

8.4/ RESULTS DISCUSSION

In Section 8.3, several assessments were made: The data was re-sampled and cleaned of outliers. The K-Mean algorithm was applied after selecting the optimal k value, resulting in three different groups of clusters. Linear regression was then performed, MAE and RMSE were calculated, breakpoints were determined, and statistics for each insee were

considered. Figures 8.5,8.6,8.7, which show the distribution of the clusters, reveal that each cluster groups the graphs that have almost the same trend and shape. To verify this conjecture, the time series decomposition (DTS) technique is applied, which envisages the trend pattern for each insee of each cluster. The DTS explained audibly that each cluster classified the insee with the same trend, as shown in Figures 8.8, 8.9 and 8.10.

Figure 8.8: Number of interventions in Cluster 0

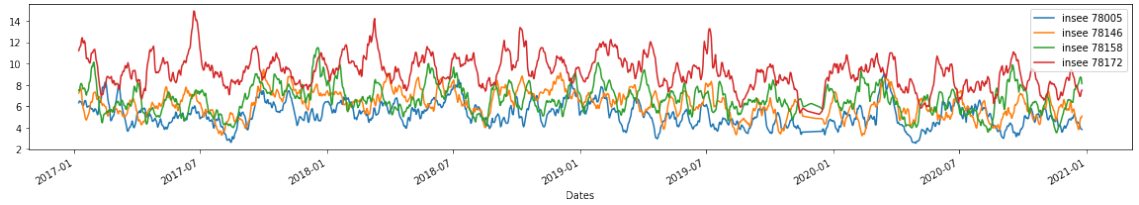


Figure 8.9: Number of interventions in Cluster 1

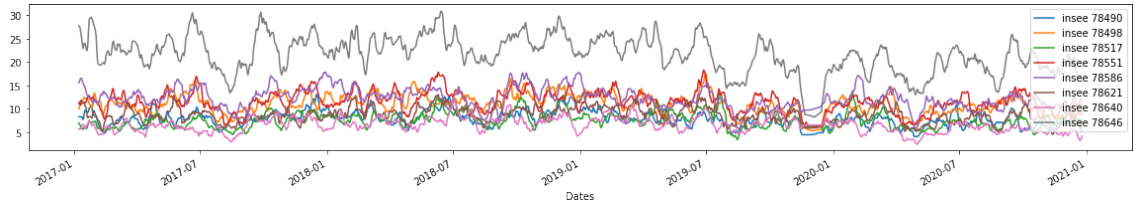
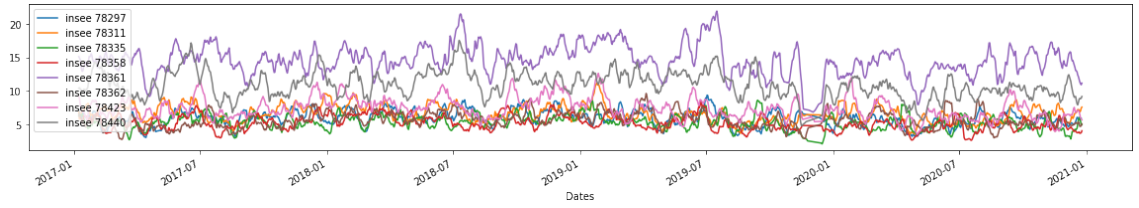


Figure 8.10: Number of interventions in Cluster 2



On the other hand, the calculation of two breakpoints for each insee didn't yield a common breakpoint as a synthesis to be inferred for cluster partitioning. In contrast, some breakpoint data (especially month and year) are duplicated in different insee, which may be relevant for other case studies.

Besides, the linear regression plots show a line of points where these points have the smallest distance to the firemen interventions of each insee. MAE and RMSE show the deviation from the actual values with an average of 3. This reflects that the dependent variables related to the number of fire department deployments do not vary much in the same insee.

Eventually, the statistics that contained information about the population size for each Insee played no role in clustering the number of fire calls. Population levels can be high or low in the same insee, which means that the number of missions can be excessive or moderate without regard to the flow of people. Nonetheless, adding population as an explanatory variable to the dataset and ignoring the time series technique may be an appropriate way forward.

8.5/ CONCLUSION

This chapter presented a case study in which the number of interventions in the Île-de-France region, particularly in Yvelines, was divided into different clusters. First, the reason why only the 20 largest insee in Yvelines were selected in terms of the number of firemen brigades was discussed. Also, the outliers have been removed from some insee that have anomalies, which are represented in the graph as large and unique peaks.
Second, the k-mean clustering technique was executed after choosing the optimal parameter k for the number of clusters. Third, a linear regression algorithm was implemented for each insee to examine the variability in the number of fire calls within the same insee. Finally, the partitioning of the clusters was analyzed to inspect which criteria were responsible for the partitioning. Breakpoints and statistics on population and jobs did not reveal a dominant conclusion.
Ultimately, each cluster contains the insee that has approximately the same number of fire calls, represented under a homogeneous trend in the DTS, without neglecting the main feature of this study, which is the time series, containing only as variables the date, the insee and the number of interventions.

IV

CONCLUSION & PERSPECTIVES

9

CONCLUSION & PERSPECTIVES

9.1/ CONCLUSION

This thesis proposed several approaches to optimize firefighting operations in order to reduce financial and material resources and improve emergency response efficiency in terms of time and quality. At the beginning (Chapters 2, 3 and 4), this dissertation discusses the scientific background of machine learning, time series, emergency operations, and technical tools. The second part (Chapters 5, 6, 7, and 8) presents contributions dealing with predictive analysis of time series in different application contexts.

The first part starts with an overview of the main applications of machine learning in our daily life in different domains such as agriculture, healthcare, marketing, commerce, and many others. Then, the integration of ML specifically in emergency response (hospitals, emergency rooms, police stations, ambulances, and fire departments) is highlighted. In addition, the challenges faced by ML such as privacy, security, reliability, completeness, and interoperability, were explained.

Secondly, the characteristics of time series forecasting and the progress over the years were introduced. Then, all related works involving machine learning on firefighters were listed and finally, in Chapter 3, all ML techniques and algorithms used in all papers were explained with all the necessary statistical tools and frameworks to complete the algorithms.

Thirdly, to complete the goal of each dissertation, the implementation of time series algorithms was performed. For this purpose, the infrastructure, including libraries, packages, frameworks, hardware, and software, was discussed.

The second part of this dissertation contained several contributions. At first sight, the dataset was studied and the number of firemen missions was predicted using different machine learning algorithms: AR, MA, ARIMA, Prophet, Simple Exponential Smoothing, Holt, and Holt Winter. The dataset was explained, processed by removing outliers, decomposed to find out the trend and seasonality of fire activity over many years, and op-

timal values for each algorithm were chosen. Then, the prediction results were analyzed, showing an effective prediction for both short and long periods, proving that fire brigade activity is not a random process but predictable.

After the initial research was completed by early 2020, the COVID-19 turned the world upside down where everything was affected. All aspects of our lives have been turned upside down. The latest data set provided by the Fire and Rescue Department SDIS-25 showed that the number of times firefighters were called out during this pandemic increased. Therefore, an analysis of firemen activity to verify the accuracy of the forecast during this sensitive and surprising time was a must.

The work begins by reviewing the highest feature importance among all available attributes, including those related to COVID-19. Then, breakpoints were identified to show an apparent change in trend before and after the onset of the intensive spread of coronavirus in France beginning in August 2020. Anomalies were detected, and finally, the prediction of firemen interventions after the replacement of anomalies and the prediction of anomalies themselves were performed. This research showed promising results in predicting irregular events.

The third study considers a new dataset that includes the type of interventions for each specific date and time instead of the number of interventions. In the first approach, 14 different subsets of data are generated for each of the 14 different types of interventions, and in the second approach, the newly generated subsets of data are merged with the original data set used in previous studies to test the efficiency of adding explanatory variables. As a result, the type of interventions played a large role in the prediction error as the size of the datasets differed from one category to another, and the 14 partial datasets showed better statistical results than the merged datasets.

Finally, the Fire and Rescue Department has provided us with a completely new data set from various regions. In all the previous research, we analyzed the fire responses in Doubs-France, but the last study included Île de France. Since we have a larger number of insee and departments, the highest number of deployments belonging to Yvelines was selected and then the largest 20 insee to achieve a feasible clustering. The number of fire departments was divided into three different clusters using the k-mean algorithm. Neither breakpoints nor population statistics provided information on how the clusters diverged. Ultimately, each cluster contained approximately the same number of deployments.

We believe that the various studies in this paper will help fire and rescue departments better understand the flow of their operations and to be prepared by knowing the rough human and material resources needed for different types of operations in different regions and even in irregular events. This, of course, will reduce their financial costs.

9.2/ PERSPECTIVES

While the study of the predictability of firefighting operations has progressed well over the past five years, a certain amount of work remains to be done in order to reach maturity in this field of research, and the widespread use of such a predictive tool.

First of all, if the prediction of calls, or of any type of intervention, at the level of the whole department gives satisfactory results, this is not the case for sub-types of intervention such as emergency rescue in big cities, or chimney fires. Predictions for such sub-types can be very useful, however. For example, there are so-called “hot” neighborhoods in the two large cities of the Doubs, which for social reasons are more prone to urban violence. This violence leads to a large number of accidents, fires, etc. in a short period of time and in a very limited area, and being able to anticipate such situations would make it possible to size the guards, or even to pre-position human and material resources. It could be possible to predict this risk to a certain extent, as these incidents frequently have an origin (police blunder, defeat at a soccer match, etc.). The means of such a prediction could be natural language processing on news from the local press and on the agitation at the level of social networks, but until now everything remains to be done.

At this level, the key is to find use cases, to find situations where predictions can be both accurate and useful. For example, as we have seen in the various articles presented here, predicting any type of intervention at the geographical level of the department and for the hour to come is done very well, and the error is really minimal. But such a prediction has no practical use. On the other hand, if we change the target from “any type of intervention” to “any phone call received by the fire department”, we remain on something easily predictable, but which gains in utility. Indeed, it allows to size the call room staff, and to have more people to pick up the phone in peak periods. The usefulness must therefore guide the choices of what to predict, in consultation with the people in the business. But it cannot be the sole driver of the research, and must be coupled with feasibility. For example, predicting the risk of chimney fires is useful and feasible at the level of the department or of fairly large regions, but will never be feasible at the level of a village. Similarly, predicting the risk of illness at the street level would certainly be very useful, but can only be envisaged if we have explanatory variables that are both very precise and very localized, and with a significant history, and we risk waiting decades before having this.

One of the first works that can be carried out combining both a real usefulness and a certain feasibility concerns crisis situations associated with natural or exceptional risks. We have carried out a first work in this sense at the level of the risk of flooding in the Doubs. This risk comes from natural rivers which, due to particular climatic conditions, have a level and flow that increase. Once a certain level is reached, the rivers overflow

and the fire department faces a peak of interventions over a short period of time. These peaks are directly linked to local weather data, which is available in real time. Even better, there is a whole network of flow and height measurements of various rivers, which are also available in real time. It is obvious that these flood intervention peaks are directly and strongly correlated to these explanatory variables, and the operational interest of such predictions is also obvious. These predictions are not without difficulty, however, because they are rare events and because we have only a small history. But these predictions can only improve over time (there is another difficulty coming from the fact that the climate is changing, in these times of global warming).

Various other hazards, natural and otherwise, are also useful to predict. Examples include forest fires, agricultural fires, coastal storms, avalanche risk, and even industrial risks. These risks are not the same from one region to another, and each has its own dynamics. It is therefore necessary to find the right explanatory variables each time and to select an ad hoc model. However, some risks are much more difficult to predict than others. For example, the floods in the Doubs mentioned above are easy to predict, because there is a real inertia, a real slowness in this event: we can observe three days in advance, sufficiently upstream of the river (the Doubs), the flood to come in Besançon. Conversely, the so-called Cevennes events taking place in the mountains in the south of France lead to spectacular rises in mountain rivers in an instant, and are therefore very difficult to predict, although dangerous.

Predictions of such risks are therefore useful, but only if a certain accuracy is achieved. To do so, we would need a lot of data, especially since we are dealing with rare phenomena. However, this is the main pitfall in this problem: the departmental fire and rescue services have only recently digitized their intervention data, and we only have at best a 7-year history for the most advanced departments. For flood interventions in the Doubs, this translates into about ten events, which is clearly insufficient to achieve good predictions on various time horizons. On the other hand, the fact that the risks are not the same from one department to another prevents, to a certain extent, to compensate for the low temporal depth of the history by a geographical multiplicity from the hundred departments of France. Unless one is willing to wait for the history to grow in time, the solution is to group the data from centers with similar profiles, after a clustering step. However, this solution has yet to be implemented. However, there is a final obstacle at this level, which for the moment has no solution: the world is evolving and the dynamics of interventions are changing over time, for various reasons related to global warming, aging of the population, new epidemics such as Covid-19, etc. Also, the history does not fully reflect what is happening now.

Once solutions to these problems are found, it would be useful to broaden the focus, and extend the predictions to all preemergency transportation. In France, the latter is

shared by three entities: the fire department, private ambulances, and public ambulances of hospital emergency services. Each of them has an impact on the others. Thus, strikes in emergency departments, the closure of small hospitals, or the reduction of licenses for the private sector, lead to an overload for the departmental fire and rescue services, and taking into account the one makes it possible to better predict what will happen to the others. And taking into account the pre-sanitary sector as a whole makes it possible to have a longer-term visibility, and for example to distribute the right number of approvals for the private sector, so that the latter can relieve the public sector while being economically viable (a fair balance must be found). Finally, emergency pre-hospital transport is not only of interest in France, and the same problems are found, mutatis mutandis, in other countries. We should be able to move on to the international level, knowing how to extract what is common to each country, and what must be adapted locally.

PUBLICATIONS

SUBMITTED PAPERS

- Roxane Mallouhy, Christophe Guyeux, Chady Abou Jaoude, Abdallah Makhoul *"Major earthquake event prediction using various machine learning algorithms"*. In **International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)**. Published in December 2019.
- Roxane Mallouhy, Christophe Guyeux, Chady Abou Jaoude, Abdallah Makhoul *"Time Series Forecasting for the Number of Firefighters Interventions"*. In **International Conference on Advanced Information Networking and Applications (AINA 2021)**. Published in May 2021.
- Roxane Mallouhy, Christophe Guyeux, Chady Abou Jaoude, Abdallah Makhoul *"Forecasting the Number of Firemen Interventions Using Exponential Smoothing Methods: A Case Study"*. In **International Conference on Advanced Information Networking and Applications (AINA 2022)**. Published in April 2022.
- Roxane Mallouhy, Christophe Guyeux, Chady Abou Jaoude, Abdallah Makhoul *"Anomalies and Breakpoint Detection for a Dataset of Firefighters' Operations During the COVID-19 Period in France"*. In **World Conference on Information Systems and Technologies (WorldCIST 2022)**. Published in April 2022.
- Roxane Mallouhy, Christophe Guyeux, Chady Abou Jaoude, Abdallah Makhoul *"Machine Learning for Predicting Firefighters' Interventions Per Type of Mission"*. In **2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)**. Published in May 2022.
- Roxane Mallouhy, Christophe Guyeux, Chady Abou Jaoude, Abdallah Makhoul *"Predicting fire brigades' operations based on their type of interventions"*. In **nternational Wireless Communications and Mobile Computing Conference (IWCMC 2022)**. Published in May 2022.
- Roxane Mallouhy, Christophe Guyeux, Chady Abou Jaoude, Abdallah Makhoul *"K-mean Clustering: a case study in Yvelines, Île-de-France"*. In **International Conference on Computational Intelligence and Communication Networks (CICN 2022)**. Presented in December 2022.

BIBLIOGRAPHY

[Saba Zafar et al.,] Saba Zafar, M. K., Khan, M. I., et Nida, H. **Application of simple exponential smoothing method for temperature forecasting in two major cities of the punjab, pakistan**.

[Galton, 1886] Galton, F. (1886). **Regression towards mediocrity in hereditary stature.** *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

[Hooker, 1921] Hooker, R. H. (1921). **Forecasting the crops from the weather**. *Quarterly Journal of the Royal Meteorological Society*, 47(198):75–100.

[Brown, 1956] Brown, R. G. (1956). **Exponential smoothing for predicting demand. cambridge, mass., arthur d. little**. *Book Exponential Smoothing for Predicting Demand*.

[Cox, 1958] Cox, D. R. (1958). **The regression analysis of binary sequences**. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

[Winters, 1960] Winters, P. R. (1960). **Forecasting sales by exponentially weighted moving averages**. *Management science*, 6(3):324–342.

[MacQueen et al., 1967] MacQueen, J., et others (1967). **Some methods for classification and analysis of multivariate observations**. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, page 281–297. Oakland, CA, USA.

[George, 1970] George, E. (1970). **Box. time series analysis: forecasting and control**.

[Sakoe et al., 1978] Sakoe, H., et Chiba, S. (1978). **Dynamic programming algorithm optimization for spoken word recognition**. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

[Hartigan et al., 1979] Hartigan, J. A., et Wong, M. A. (1979). **Algorithm as 136: A k-means clustering algorithm**. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108.

[Lloyd, 1982] Lloyd, S. (1982). **Least squares quantization in pcm**. *IEEE transactions on information theory*, 28(2):129–137.

[Ledolter et al., 1984] Ledolter, J., et Abraham, B. (1984). **Some comments on the initialization of exponential smoothing**. *Journal of Forecasting*, 3(1):79–84.

[Gardner Jr, 1985] Gardner Jr, E. S. (1985). **Exponential smoothing: The state of the art**. *Journal of forecasting*, 4(1):1–28.

[Snyder, 1985] Snyder, R. (1985). **Recursive estimation of dynamic linear models**. page 272–276.

[Sweet et al., 1988] Sweet, A. L., et Wilson, J. R. (1988). **Pitfalls in simulation-based evaluation of forecast monitoring schemes**. *International Journal of Forecasting*, 4(4):573–579.

[Bartolomei et al., 1989] Bartolomei, S. M., et Sweet, A. L. (1989). **A note on a comparison of exponential smoothing methods for forecasting seasonal series**. *International Journal of Forecasting*, 5(1):111–116.

[Hansen, 1990] Hansen, P. C. (1990). **Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank**. *SIAM Journal on Scientific and Statistical Computing*, 11(3):503–518.

[Gardner Jr, 1993] Gardner Jr, E. S. (1993). **Forecasting the failure of component parts in computer systems: A case study**. *International Journal of Forecasting*, 9(2):245–253.

[Miller et al., 1993] Miller, T., et Liberatore, M. (1993). **Seasonal exponential smoothing with damped trends: An application for production planning**. *International Journal of Forecasting*, 9(4):509–515.

[Rosas et al., 1994] Rosas, A. L., et Guerrero, V. M. (1994). **Restricted forecasts using exponential smoothing techniques**. *International Journal of Forecasting*, 10(4):515–527.

[Cortes et al., 1995] Cortes, C., et Vapnik, V. (1995). **Support-vector networks**. *Machine learning*, 20(3):273–297.

[Zhang et al., 1996] Zhang, T., Ramakrishnan, R., et Livny, M. (1996). **Birch: an efficient data clustering method for very large databases**. *ACM sigmod record*, 25(2):103–114.

[Kavallieratos et al., 1997] Kavallieratos, E., Antoniades, N., Fakotakis, N., et Kokkinakis, G. (1997). **Extraction and recognition of handwritten alphanumeric characters from application forms**. In *Proceedings of 13th International Conference on Digital Signal Processing*, page 695–698. IEEE.

[Hochreiter et al., 1997] Hochreiter, S., et Schmidhuber, J. (1997). **Long short-term memory**. *Neural computation*, 9(8):1735–1780.

[Klein et al., 1997] Klein, J. L., et Klein, D. (1997). **Statistical visions in time: a history of time series analysis, 1662-1938**. Cambridge University Press.

[Balakrishnama et al., 1998] Balakrishnama, S., et Ganapathiraju, A. (1998). **Linear discriminant analysis-a brief tutorial**. *Institute for Signal and information Processing*, 18(1998):1–8.

[Williams et al., 1999] Williams, D. W., et Miller, D. (1999). **Level-adjusted exponential smoothing for modeling planned discontinuities**. *International Journal of Forecasting*, 15(3):273–289.

[Werts et al., 2000] Werts, N., et Adya, M. (2000). **Data mining in healthcare: issues and a research agenda**. page 98.

[Tibshirani et al., 2001] Tibshirani, R., Walther, G., et Hastie, T. (2001). **Estimating the number of clusters in a data set via the gap statistic**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

[Fernandes, 2001] Fernandes, P. A. M. (2001). **Fire spread prediction in shrub fuels in portugal**. *Forest ecology and management*, 144(1-3):67–74.

[Friedman, 2001] Friedman, J. H. (2001). **Greedy function approximation: a gradient boosting machine**. page 1189–1232.

[Grubb et al., 2001] Grubb, H., et Mason, A. (2001). **Long lead-time forecasting of uk air passengers by holt–winters methods with damped trend**. *International Journal of Forecasting*, 17(1):71–82.

[Hyndman et al., 2002] Hyndman, R. J., Koehler, A. B., Snyder, R. D., et Grose, S. (2002). **A state space framework for automatic forecasting using exponential smoothing methods**. *International Journal of forecasting*, 18(3):439–454.

[Taylor, 2003] Taylor, J. W. (2003). **Exponential smoothing with a damped multiplicative trend**. *International journal of Forecasting*, 19(4):715–725.

[Kargupta et al., 2003] Kargupta, H., Datta, S., Wang, Q., et Sivakumar, K. (2003). **On the privacy preserving properties of random data perturbation techniques**. page 99–106. IEEE.

[Muralidhar et al., 2003] Muralidhar, K., et Sarathy, R. (2003). **A theoretical basis for perturbation methods**. *Statistics and Computing*, 13(4):329–335.

[Oliveira et al., 2004] Oliveira, S., et Zaiane, O. (2004). **Data perturbation by rotation for privacy-preserving clustering**.

[Holt, 2004] Holt, C. C. (2004). **Forecasting seasonals and trends by exponentially weighted moving averages**. *International journal of forecasting*, 20(1):5–10.

[Segal, 2004] Segal, M. R. (2004). **Machine learning benchmarks and random forest regression**.

[Vaidya et al., 2004] Vaidya, J., et Clifton, C. (2004). **Privacy-preserving data mining: Why, how, and when**. *IEEE Security & Privacy*, 2(6):19–27.

[Verykios et al., 2004] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., et Theodoridis, Y. (2004). **State-of-the-art in privacy preserving data mining**. *ACM Sigmod Record*, 33(1):50–57.

[Özgür, 2005] Özgür, K. (2005). **Daily river flow forecasting using artificial neural networks and auto-regressive models**. *Turkish Journal of Engineering and Environmental Sciences*, 29(1):9–20.

[Geurts et al., 2006] Geurts, P., Ernst, D., et Wehenkel, L. (2006). **Extremely randomized trees**. *Machine learning*, 63(1):3–42.

[Lai et al., 2006] Lai, K. K., Yu, L., Wang, S., et Huang, W. (2006). **Hybridizing exponential smoothing and neural network for financial time series predication**. page 493–500. Springer.

[Shen et al., 2007] Shen, A., Tong, R., et Deng, Y. (2007). **Application of classification models on credit card fraud detection**. page 1–4. IEEE.

[Youn et al., 2007] Youn, S., et McLeod, D. (2007). **A comparative study for email classification**. page 387–391. Springer.

[Kuhn et al., 2007] Kuhn, A., Ducasse, S., et Gîrba, T. (2007). **Semantic clustering: Identifying topics in source code**. *Information and software technology*, 49(3):230–243.

[Guyon et al., 2008] Guyon, I., Gunn, S., Nikravesh, M., et Zadeh, L. A. (2008). **Feature extraction: foundations and applications**, volume 207. Springer.

[Chen et al., 2008] Chen, P., Yuan, H., et Shu, X. (2008). **Forecasting crime using the arima model**. In *2008 fifth international conference on fuzzy systems and knowledge discovery*, page 627–630. IEEE.

[Jones et al., 2008] Jones, S. S., Thomas, A., Evans, R. S., Welch, S. J., Haug, P. J., et Snow, G. L. (2008). **Forecasting daily patient volumes in the emergency department**. *Academic Emergency Medicine*, 15(2):159–170.

[Cunningham et al., 2008] Cunningham, P., Cord, M., et Delany, S. J. (2008). **Supervised learning**. page 21–49. Springer.

[Wu et al., 2008] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et others (2008). **Top 10 algorithms in data mining**. *Knowledge and information systems*, 14(1):1–37.

[Hosseinkhah et al., 2009] Hosseinkhah, F., Ashktorab, H., Veen, R., et others (2009). **Challenges in data mining on medical databases**. page 1393–1404. IGI global.

[Zien et al., 2009] Zien, A., Krämer, N., Sonnenburg, S., et Rätsch, G. (2009). **The feature importance ranking measure**. page 694–709. Springer.

[Peterson, 2009] Peterson, L. E. (2009). **K-nearest neighbor**. *Scholarpedia*, 4(2):1883.

[Yan et al., 2009] Yan, X., et Su, X. (2009). **Linear regression analysis: theory and computing**. World Scientific.

[Bradstock et al., 2009] Bradstock, R. A., Cohn, J., Gill, A. M., Bedward, M., et Lucas, C. (2009). **Prediction of the probability of large fires in the sydney region of south-eastern australia using fire weather**. *International Journal of Wildland Fire*, 18(8):932–943.

[Webb et al., 2010] Webb, G. I., Keogh, E., et Miikkulainen, R. (2010). **Naïve bayes.** *Encyclopedia of machine learning*, 15:713–714.

[Ip et al., 2010] Ip, A., Fong, S., et Liu, E. (2010). **Optimization for allocating bev recharging stations in urban areas by using hierarchical clustering**. page 460-465.

[Nayak et al., 2011] Nayak, G., et Devi, S. (2011). **A survey on privacy preserving data mining: approaches and techniques**. *International Journal of Engineering Science and Technology*, 3(3):2127–2133.

[Saha, 2011] Saha, D. (2011). **Web text classification using a neural network**. page 57–60. IEEE.

[Yue et al., 2012] Yue, Y., Marla, L., et Krishnan, R. (2012). **An efficient simulation-based approach to ambulance fleet allocation and dynamic redeployment**. In *Proceedings of the AAAI Conference on Artificial Intelligence*, page 398–405.

[Ostertagová, 2012] Ostertagová, E. (2012). **Modelling using polynomial regression**. *Procedia Engineering*, 48:500–506.

[Devi et al., 2013] Devi, B. U., Sundar, D., et Alli, P. (2013). **An effective time series analysis for stock trend prediction using arima model for nifty midcap-50**. *International Journal of Data Mining & Knowledge Management Process*, 3(1):65.

[Kaur et al., 2013] Kaur, D., et Jyoti, K. (2013). **Enhancement in the performance of k-means algorithm**. *International Journal of Computer Science and Communication Engineering*, 2(1):29–32.

[Gang et al., 2013] Gang, T.-T., Yang, J., et Zhao, Y. (2013). **Multivariate control chart based on the highest possibility region**. *Journal of Applied Statistics*, 40(8):1673–1681.

[Sharma et al., 2013] Sharma, B. R., Kaur, D., et Manju, A. (2013). **Review on data mining: its challenges, issues and applications**. *International Journal of Current Engineering and Technology*, 3(2):695–700.

[Tomar et al., 2013] Tomar, D., et Agarwal, S. (2013). **A survey on data mining approaches for healthcare**. *International Journal of Bio-Science and Bio-Technology*, 5(5):241–266.

[Chujai et al., 2013] Chujai, P., Kerdprasop, N., et Kerdprasop, K. (2013). **Time series analysis of household electric consumption with arima and arma models**. In *Proceedings of the international multiconference of engineers and computer scientists*, page 295–300. IAENG Hong Kong.

[Kumar et al., 2014] Kumar, M., et Anand, M. (2014). **An application of time series arima forecasting model for predicting sugarcane production in india**. *Studies in Business and Economics*, 9(1):81–94.

[Ali et al., 2014] Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., et Faraj, R. H. (2014). **Data normalization and standardization: a technical report**. *Mach Learn Tech Rep*, 1(1):1–6.

[Taherdoost et al., 2014] Taherdoost, H., Sahibuddin, S., et Jalaliyoon, N. (2014). **Exploratory factor analysis; concepts and theory**.

[Perlich et al., 2014] Perlich, C., Dalessandro, B., Raeder, T., Stitelman, O., et Provost, F. (2014). **Machine learning for targeted display advertising: Transfer learning in action**. *Machine learning*, 95(1):103–127.

[Chen et al., 2014] Chen, S., Lan, X., Hu, Y., Liu, Q., et Deng, Y. (2014). **The time series forecasting: from the aspect of network**. *arXiv preprint arXiv:1403.1713*.

[Ohno-Machado et al., 2015] Ohno-Machado, L., et in Chief, E. (2015). **Mining electronic health record data: finding the gold nuggets**. *Journal of the American Medical Informatics Association*, 22(5):937–937.

[Anggrainingsih et al., 2015] Anggrainingsih, R., Aprianto, G. R., et Sihwi, S. W. (2015). **Time series forecasting using exponential smoothing to predict the number of website visitor of sebelas maret university**. page 14–19. IEEE.

[Chen et al., 2015] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., et others (2015). **Xgboost: extreme gradient boosting**. *R package version 0.4-2*, 1(4):1–4.

[Jolliffe et al., 2016] Jolliffe, I. T., et Cadima, J. (2016). **Principal component analysis: a review and recent developments**. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

[Denaxas et al., 2016] Denaxas, S. C., Asselbergs, F. W., et Moore, J. H. (2016). **The tip of the iceberg: challenges of accessing hospital electronic health record data for biological data mining**.

[Celebi et al., 2016] Celebi, M. E., et Aydin, K. (2016). **Unsupervised learning algorithms**. Springer.

[Chen et al., 2016] Chen, T., et Guestrin, C. (2016). **XGBoost: A scalable tree boosting system**. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794, New York, NY, USA. ACM.

[Jadhav et al., 2017] Jadhav, V., CHINNAPPA, R. B., et Gaddi, G. (2017). **Application of arima model for forecasting agricultural prices**.

[O'Connor et al., 2017] O'Connor, C. D., Calkin, D. E., et Thompson, M. P. (2017). **An empirical machine learning method for predicting potential fire control locations for pre-fire planning and operational fire management**. *International journal of wildland fire*, 26(7):587–597.

[Li et al., 2017] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., et Liu, H. (2017). **Feature selection: A data perspective**. *ACM computing surveys (CSUR)*, 50(6):1–45.

[Zhang et al., 2017] Zhang, H., Zhang, S., Wang, P., Qin, Y., et Wang, H. (2017). **Forecasting of particulate matter time series using wavelet analysis and wavelet-arma/arima model in taiyuan, china**. *Journal of the Air & Waste Management Association*, 67(7):776–788.

[Bansal et al., 2017] Bansal, A., Sharma, M., et Goel, S. (2017). **Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining**. *International Journal of Computer Applications*, 157(6):0975–8887.

[Ke et al., 2017] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., et Liu, T.-Y. (2017). **Lightgbm: A highly efficient gradient boosting decision tree**. *Advances in neural information processing systems*, 30:3146–3154.

[Blobel et al., 2018] Blobel, B., Oeming, F., et Ruotsalainen, P. (2018). **Data modeling challenges of advanced interoperability**. *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth: Proceedings of MIE 2018*.

[Marutho et al., 2018] Marutho, D., Hendra Handaka, S., Wijaya, E., et Muljono (2018). **The determination of cluster number at k-mean using elbow method and purity evaluation on headline news**. page 533-538.

[Taylor et al., 2018] Taylor, S. J., et Letham, B. (2018). **Forecasting at scale**. *The American Statistician*, 72(1):37–45.

[Graham et al., 2018] Graham, B., Bond, R., Quinn, M., et Mulvenna, M. (2018). **Using data mining to predict hospital admissions from the emergency department**. *IEEE Access*, 6:10458–10469.

[Couchot et al., 2019] Couchot, J.-F., Guyeux, C., et Royer, G. (2019). **Anonymously forecasting the number and nature of firefighting operations**. page 1–8.

[Tamba et al., 2019] Tamba, S. P., Batubara, M. D., Purba, W., Sihombing, M., Siregar, V. M. M., et Banjarnahor, J. (2019). **Book data grouping in libraries using the k-means clustering method**. In *Journal of Physics: Conference Series*, page 012074. IOP Publishing.

[Hahsler et al., 2019] Hahsler, M., Piekenbrock, M., et Doran, D. (2019). **dbscan: Fast density-based clustering with R**. *Journal of Statistical Software*, 91(1):1–30.

[Guyeux et al., 2019] Guyeux, C., Nicod, J.-M., Varnier, C., Al Masry, Z., Zerhouni, N., Omri, N., et Royer, G. (2019). **Firemen prediction by using neural networks: a real case study**. page 541 - 552, London, United Kingdom.

[eps, 2019] (2019). **French firefighters demand more staffing, higher pay, and better possibilities for career development**.

[Singh et al., 2019] Singh, K., Shastri, S., Bhadwal, A. S., Kour, P., Kumari, M., Sharma, A., et Mansotra, V. (2019). **Implementation of exponential smoothing for forecasting time series data**. *Int. J. Sci. Res. Comput. Sci. Appl. Manag. Stud*, 8.

[Ñahuis et al., 2019] Ñahuis, S. L. C., Guyeux, C., Arcolezi, H. H., Couturier, R., Royer, G., et Lotufo, A. D. P. (2019). **Long short-term memory for predicting firemen interventions**. page 1132-1137.

[Coffield et al., 2019] Coffield, S. R., Graff, C. A., Chen, Y., Smyth, P., Foufoula-Georgiou, E., et Randerson, J. T. (2019). **Machine learning to predict final fire size at the time of ignition**. *International journal of wildland fire*, 28(11):861–873.

[Ahani et al., 2019] Ahani, A., Nilashi, M., Ibrahim, O., Sanzogni, L., et Weaven, S. (2019). **Market segmentation and travel choice prediction in spa hotels through tripadvisor's online reviews**. *International Journal of Hospitality Management*, 80:52–77.

[Akiba et al., 2019] Akiba, T., Sano, S., Yanase, T., Ohta, T., et Koyama, M. (2019). **Optuna: A next-generation hyperparameter optimization framework**. page 2623–2631.

[Pirklbauer et al., 2019] Pirklbauer, K., et Findling, R. D. (2019). **Predicting the category of fire department operations**. page 659–663.

[Lian et al., 2019] Lian, X., Melancon, S., Presta, J.-R., Reevesman, A., Spiering, B., et Woodbridge, D. (2019). **Scalable real-time prediction and analysis of san francisco fire department response times**. page 694–699. IEEE.

[Lengyel et al., 2019] Lengyel, A., et Botta-Dukát, Z. (2019). **Silhouette width using generalized mean—a flexible method for assessing clustering efficiency**. *Ecology and evolution*, 9(23):13231–13243.

[Gotschaux, 2019] Gotschaux, F. (2019). **Trop d'interventions de secours aux personnes : Les pompiers du calvados en appellent au premier ministre**.

[Okamoto Jr et al., 2020] Okamoto Jr, J., Roque, A. C., Schiavo, F., Sguerra, B. M., Miyamoto, B. A., Mourão, F. A., Alves, T. Y., et de Paula Suiti, A. (2020). **Addressing the golden hour: A machine learning approach to improve emergency response time**.

[Kang et al., 2020] Kang, D.-Y., Cho, K.-J., Kwon, O., Kwon, J.-m., Jeon, K.-H., Park, H., Lee, Y., Park, J., et Oh, B.-H. (2020). **Artificial intelligence algorithm to predict the need for critical care in prehospital emergency medical services**. *Scandinavian journal of trauma, resuscitation and emergency medicine*, 28(1):1–8.

[Keshavarzi Arshadi et al., 2020] Keshavarzi Arshadi, A., Webb, J., Salem, M., Cruz, E., Calad-Thomson, S., Ghadirian, N., Collins, J., Diez-Cecilia, E., Kelly, B., Goodarzi, H., et others (2020). **Artificial intelligence for covid-19 drug discovery and vaccine development**. *Frontiers in Artificial Intelligence*, 3.

[Cerna et al., 2020] Cerna, S., Guyeux, C., Arcolezi, H. H., et Royer, G. (2020). **Boosting methods for predicting firemen interventions**. page 001–006. IEEE.

[Cerna et al., 2020] Cerna, S., Guyeux, C., Arcolezi, H. H., Couturier, R., et Royer, G. (2020). **A comparison of lstm and xgboost for predicting firemen interventions**. page 424–434. Springer.

[Pinter et al., 2020] Pinter, G., Felde, I., Mosavi, A., Ghamisi, P., et Gloaguen, R. (2020). **Covid-19 pandemic prediction for hungary; a hybrid machine learning approach**. *Mathematics*, 8(6):890.

[Brownlee, 2020] Brownlee, J. (2020). **Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python**. Machine Learning Mastery.

[Redfield et al., 2020] Redfield, C., Tlimat, A., Halpern, Y., Schoenfeld, D. W., Ullman, E., Sontag, D. A., Nathanson, L. A., et Horng, S. (2020). **Derivation and validation of a machine learning record linkage algorithm between emergency medical services and the emergency department**. *Journal of the American Medical Informatics Association*, 27(1):147–153.

[Fu et al., 2020] Fu, W., et Perry, P. O. (2020). **Estimating the number of clusters using cross-validation**. *Journal of Computational and Graphical Statistics*, 29(1):162–173.

[Arcolezi et al., 2020] Arcolezi, H. H., Couchot, J.-F., Cerna, S., Guyeux, C., Royer, G., Al Bouna, B., et Xiao, X. (2020). **Forecasting the number of firefighter interventions per region with local-differential-privacy-based data**. *Computers & Security*, 96:101888.

[Raveendran, 2020] Raveendran, N. (2020). **Future of smart firefighting with artificial intelligence**.

[Miller et al., 2020] Miller, J. M., Cullingham, C. I., et Peery, R. M. (2020). **The influence of a priori grouping on inference of genetic clusters: simulation study and literature review of the dapc method**. *Heredity*, 125(5):269–280.

[col, 2020] (2020). **Les collectivités locales en chiffres 2020 - collectivites-locales.gouv.fr**.

[Sujath et al., 2020] Sujath, R. a. a., Chatterjee, J. M., et Hassanien, A. E. (2020). **A machine learning forecasting model for covid-19 pandemic in india**. *Stochastic Environmental Research and Risk Assessment*, 34(7):959–972.

[Martínez et al., 2020] Martínez, A., Schmuck, C., Pereverzyev Jr, S., Pirker, C., et Haltmeier, M. (2020). **A machine learning framework for customer purchase prediction in the non-contractual setting**. *European Journal of Operational Research*, 281(3):588–596.

[Argawu, 2020] Argawu, A. (2020). **Modeling and forecasting of covid-19 new cases in the top 10 infected african countries using regression and time series models**. *medRxiv*.

[Zhang et al., 2020] Zhang, T., Wu, Q., et Zhang, Z. (2020). **Pangolin homology associated with 2019-ncov**. *BioRxiv*.

[Cerna et al., 2020] Cerna, S., Guyeux, C., Royer, G., Chevallier, C., et Plumerel, G. (2020). **Predicting fire brigades operational breakdowns: A real case study**. *Mathematics*, 8(8):1383.

[Ali, 2020] Ali, M. (2020). **Pycaret: An open source, low-code machine learning library in python**. *PyCaret version*, 2.

[Truong et al., 2020] Truong, C., Oudre, L., et Vayatis, N. (2020). **Selective review of offline change point detection methods**. *Signal Processing*, 167:107299.

[Voyant et al., 2020] Voyant, C., Notton, G., Duchaud, J.-L., Almorox, J., et Yaseen, Z. M. (2020). **Solar irradiation prediction intervals based on box–cox transformation and univariate representation of periodic autoregressive model**. *Renewable Energy Focus*, 33:43–53.

[Yasar et al., 2020] Yasar, H., et Kilimci, Z. H. (2020). **Us dollar/turkish lira exchange rate forecasting model based on deep learning methodologies and time series analysis**. *Symmetry*, 12(9):1553.

[Kassania et al., 2021] Kassania, S. H., Kassanib, P. H., Wesolowskic, M. J., Schneidera, K. A., et Detersa, R. (2021). **Automatic detection of coronavirus disease (covid-19) in x-ray and ct images: a machine learning based approach**. *Biocybernetics and Biomedical Engineering*, 41(3):867–879.

[Fang et al., 2021] Fang, H., Lo, S., Zhang, Y., et Shen, Y. (2021). **Development of a machine-learning approach for identifying the stages of fire development in residential room fires**. *Fire Safety Journal*, 126:103469.

[Cerna et al., 2021] Cerna, S., Arcolezi, H. H., Guyeux, C., Royer-Fey, G., et Chevallier, C. (2021). **Machine learning-based forecasting of firemen ambulances' turnaround time in hospitals, considering the covid-19 impact**. *Applied soft computing*, 109:107561.

[Ho et al., 2021] Ho, W. K., Tang, B.-S., et Wong, S. W. (2021). **Predicting property prices with machine learning algorithms**. *Journal of Property Research*, 38(1):48–70.

[Raiter, 2021] Raiter, O. (2021). **Segmentation of bank consumers for artificial intelligence marketing**. *International Journal of Contemporary Financial Issues*, 1(1):39–54.

[Emioma et al., 2021] Emioma, C., et Edeki, S. (2021). **Stock price prediction using machine learning on least-squares linear regression basis**. In *Journal of Physics: Conference Series*, page 012058. IOP Publishing.

[Sujan et al., 2022] Sujan, M., Thimbleby, H., Habli, I., Cleve, A., Maaløe, L., et Rees, N. (2022). **Assuring safe artificial intelligence in critical ambulance service response: study protocol**. *British Paramedic Journal*, 7(1):36–42.

[pom, 2022] (2022). **Chiffres clés**.

[Guyeux et al., 2022] Guyeux, C., Makhoul, A., et Bahi, J. (2022). **How to build an optimal and operational knowledge base to predict firefighters' interventions**. In *Intelligent Systems Conference (IntelliSys 2022)*, Amsterdam, Netherlands.

[Schlager et al., 2022] Schlager, T., Christen, M., et others (2022). **Market segmentation**. page 939.

[Angelini et al., 2022] Angelini, F., Widera, P., Mobasheri, A., Blair, J., Struglics, A., Uebelhoer, M., Henrotin, Y., Marijnissen, A. C., Kloppenburg, M., Blanco, F. J., et others (2022). **Osteoarthritis endotype discovery via clustering of biochemical marker data**. *Annals of the Rheumatic Diseases*, 81(5):666–675.

List of Figures

List of Tables

Document generated with LaTeX and:
the LaTeX style for PhD Thesis created by S. Galland — http://www.multiagent.fr/ThesisStyle
the tex-upmethodology package suite — http://www.arakhne.org/tex-upmethodology/

école doctorale **sciences pour l'ingénieur et microtechniques**

Title: Predictive analysis of time series in various application contexts

Keywords: Time series Forecasting, Clustering, Data mining, Machine learning, Firefighters' interventions, Feature selection, Breakpoint, Anomalies detection, Statistical Features.

Abstract:

Emergency medical transport in France is triggered by the dispatch of an ambulance, either by the SMUR, SAMU, by a private ambulance company, or by the fire department after dialing one of the emergency numbers. Since accidents are related to human activities, which in turn depend on the time of day, season, weather, climate, special events, etc., the emergency response is not hazardous but predictable. The flows of many actors are predictable to some extent, especially because of their seasonality. Being able to predict such operations makes it possible to put in place strategies for planning that could be very helpful in managing the emergency sector, which is currently in crisis. Forecasting firefighter interventions for the short term allows for better planning for paramedic leave at the emergency level, while forecasting for the long term facilitates planning of future human and material resources. In this context, the collection of data from different streams varies over periods ranging from a few years to twenty years, the aim being to use these different flows both to analyze their dynamics and to make more or less long-term forecasts. Some of these data streams have already been used in a supervised learning approach that requires the continuous collection of a set of explanatory variables, which proves to be complex for an operational device: scripts need to be set up to retrieve these variables on an hourly basis, scheduled periodically for new machine learning, etc. As a result, different approaches have been studied and applied to different firefighter datasets provided by the fire and rescue department, with the main objective being to study such operations for better future planning and management of the emergency response at lower complexity.

Titre : Analyse de flux par des techniques de séries temporelles

Mots-clés : Séries temporelles, Clustering, Data mining, Machine learning, Interventions des sapeurs-pompiers, Sélection des attributs, Breakpoint, Détection d'anomalies, Paramètres statistiques.

Résumé :

Le transport sanitaire d'urgence est enclenché, en France, suite à l'appel à un des numéros d'urgence, et suite à cet appel une ambulance est envoyée. Les accidents étant liés à l'activité humaine, qui elle-même est conditionnée à l'heure dans le jour, à la saison, au temps qu'il fait, etc., la sollicitation pour du secours à personnes n'est donc pas aléatoire. Les flux des differents opérateurs sont prévisibles, dans une certaine mesure, notamment du fait de leur caractère saisonnier. Et parvenir à les prévoir rend possible la mise en place de stratégies de planifications, qui pourraient aider grandement à la gestion de ce secteur actuellement en crise. Par exemple, être en mesure de prévoir la sollicitation à l'horizon de quelques heures, chez les pompiers, leur permet d'anticiper le besoin en pompiers volontaires. Avoir une visibilité à court terme permet de planifier au mieux les congés des ambulanciers ou au niveau des urgences, quand une visibilité à long terme aide à la planification des besoins futurs, tant matériel qu'humain. Dans ce contexte, la collecte de données de différentes filières varie sur des périodes s'étalant de quelques à une vingtaine d'années. L'objectif consiste à exploiter au mieux ces flux, tant pour en analyser la dynamique que pour être en mesure d'effectuer des prévisions à plus ou moins long terme. Certains de ces flux ont d'ores et déjà été exploités dans une approche d'apprentissage supervisé, qui nécessite la collecte en continu d'un certain nombre de variables explicatives, ce qui s'avère complexe à mettre en oeuvre pour un dispositif opérationnel: des scripts doivent être mis en place pour récupérer à chaque heure ces variables, planifier périodiquement de nouveaux apprentissages automatiques, etc. En conséquence, différentes approches ont été appliquées sur différents jeux de données de pompiers fournis par le service départemental d'incendie et de secours, avec l'objectif principal d'établir une meilleure planification et une meilleure gestion future des pompiers à moindre complexité.

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon, France