K-mean Clustering: a case study in Yvelines, Île-de-France

Roxane Elias Mallouhy
Prince Mohammad Bin Fahd University
Al Khobar, Kingdom of Saudi Arabia
reliasmallouhy@pmu.edu.sa

Christophe Guyeux
Univ. Bourgogne Franche-Comté,
Belfort, France
christophe.guyeux@univ-fcomte.fr

Chady Abou Jaoude
Antonine University,
Baabda, Lebanon
chady.aboujaoude@ua.edu.lb

Abdallah Makhoul
Univ. Bourgogne Franche-Comté,
Belfort, France
abdallah.makhoul@univ-fcomte.fr

***Abstract*—Cluster analysis is widely used in various fields to classify data that have similarities into 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 service 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. In this paper, a collection of fire operations in different departments of the Île-de-France was made, but only the information on the 20 largest departments of Yvelines was selected. After choosing the optimal k-value, the K-means clustering method was applied, and further research was carried out to summarize the criteria by which the Insee were classified into clusters. Breakpoints were detected, statistics were obtained for each insee, linear regression was implemented, and finally time series decomposition was performed. The results show that the clusters are well separated and have a dense grouping of insee with approximately the same trend, meaning the same number of interventions.**

***Index Terms*—Data mining, time series, clustering, k-mean**

I. Introduction

Data mining is the process of dealing 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, baking, fraud detection, cybersecurity, risk management, mathematics, education, and medicines. It is also used in social networking, robotics, signal processing, computer vision, mobile computing and many more. Data mining is mainly used to make better/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.
In almost every field, investigations are conducted over time. Numerical measurements of records made at regular intervals is called time series, such as, annually (e.g., number of student enrollment [1]), quarterly (e.g., tourist flow [2]), monthly (e.g., retail sales [3]), weekly (e.g., patient visit to emergency department [4]), daily (e.g., bitcoin prices [5]), or hourly (e.g., weather temperature [6]).

In short, time series mining is a phase of discovering data patterns to make time series forecasting for long or short periods of time. The main components that distinguish time series from other machine learning techniques are:

- trend: upward, downward or stationary movement of the slope
- seasonal: repeated pattern within a given period
- irregular: unpredictable and uncontrollable fluctuations

Given the vast amount of data generated in healthcare, machine learning has played an important role in recent years. Improving the medical sector relies heavily on the accurate and successful application of time series data. One of the areas of healthcare where time series can be applied is fire calls, as the number of missions is directly related to human activities, climate, weather, holidays, and many other events. Therefore,

it is judicious to assume that accidents tend to occur during the day rather than at night, when people are driving, going to work, and using private or public transportation. In addition, human activity is much significant on New Year's Eve than on a normal day when most citizens are resting at home at midnight. In addition, fires are more likely to occur in forests in the summer because of the higher temperatures than in the winter, when rain and snow prevail. To put it in a nutshell: Fire operations are not considered hazardous events, hence, machine learning is an efficient technique that can be applied.

The ability to predict the number of firefighters' missions will reduce material and human resources, thus playing a positive role in the health sector, which is currently facing an economic crisis due to the closure of small hospitals and budget cuts. Moreover, firefighters in France are not only called to put out fires, but they are also the first to transport people to/from hospitals by vehicles or helicopters, they take care of the very elderly especially during the Covid-19 period, and they can help with births, drownings, suicides, and many other emergencies. About 250000 firefighters, called 'saveur de pompier' in France, are well trained to enhance their skills in different situations.

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 for the fire departments in the region Île-de-France. The remainder of this paper is organized as follows: Section II summarizes related work on general clustering and specifically studies on the dataset of firemen operations, while Section III presents the materials and methods used in the experiments. Furthermore, Section IV discusses the clustering results and Section V draws a conclusion.

II. Related work

Cluster analysis, which aims to discover groups in data, is used in numerous applications, including marketing, biology, geology, library, 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 [7] 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 [8], 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, [9] 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 bring more books that are interesting to readers.
In the field of urban planning, [10] 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 grid, 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 [11], linguistic information from source code such as comments and identifier names are 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. 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 conducted the dataset of the Doubs region, France. The study began in mid-2019 when [12] 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, [13] used extreme gradient boosting on anonymous data to predict the number and type of firefighters' deployments by civil protection services. Also, in [14] 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 [15], 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 [16], 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, [17] 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 re-

sponses.

Two articles also conducted a study of firemen interventions during the Covid-19 period: In [18], 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. [19] detected breakpoints caused by the Covid 19 pandemic using the XGBoost machine learning algorithm, and feature importance was calculated before and after such a rare event.

Four different ML algorithms (Autoregression, Moving Average, Autoregressive Integrated Moving Average, and Prophet) were applied in [20] to examine the best performance in predicting fire operations for long or short time periods, as well as exponential smoothing methods in [21]. In further research, not only the number of firefighters was predicted, but also the type of operations such as births, fires, emergency human assistance, and many other [22], [23] were considered. Finally, [24] 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.

III. Materials and methods

A. 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 interventions. 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 on Île-de-France.

Table I shows the total number of insee and interventions grouped by department. As can be observed, there is a large discrepancy between the number of interventions in the different departments, so performing clustering experiments for the entire dataset does not yield a convincing result. Thereby, 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 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 II.

TABLE I
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

TABLE II
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

B. Dictionaries

After cleaning and preparing the repository, a data structure called 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 2, the number of items that make up each dictionary is almost in the same range, which

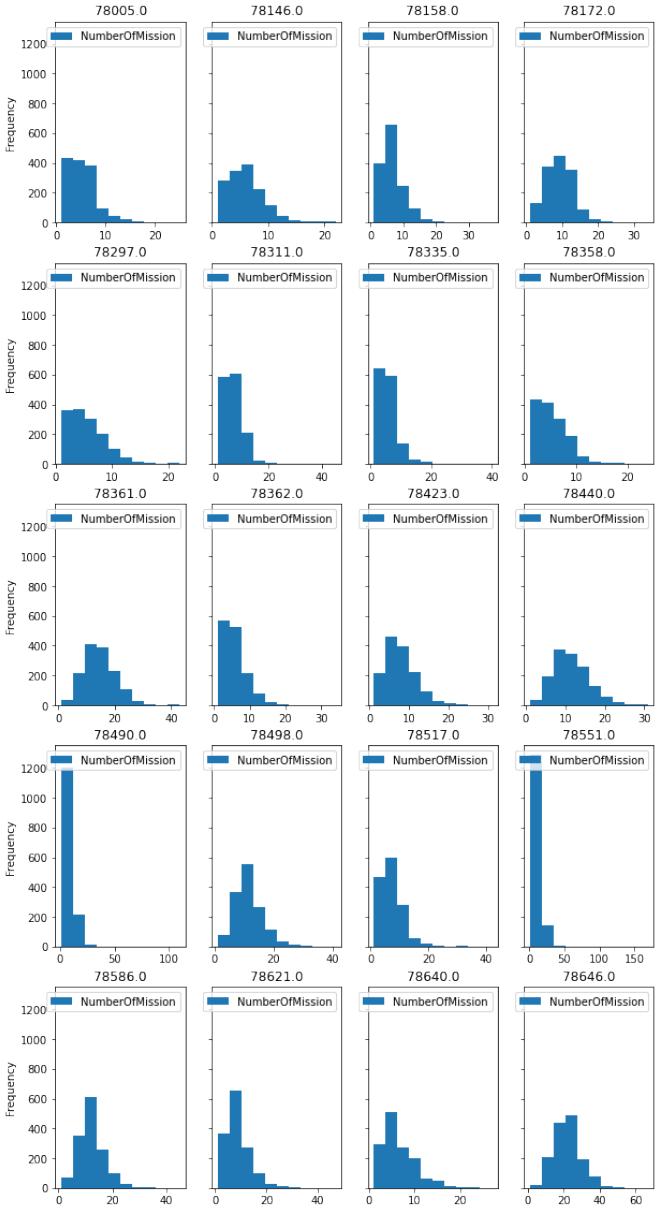

Fig. 1. Frequency of each number of missions for all insee of Yvelines

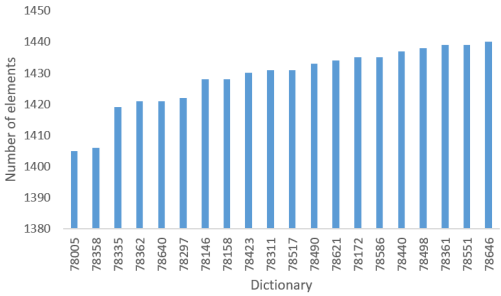

Fig. 2. Number of elements of each dictionary

makes the clustering procedure more practical than keeping all departments and all insee in the experimentation.

C. Clustering technique

1) Overview: Clustering is a multivariate data mining analysis method that uses distance measurements to divide objects into homogeneous disjoint 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 resemblance 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 [25], [26].

Particularly, 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 3, displaying visibly 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 of the corresponding dictionary. In consequence, both insee 78551 and 78490 have been changed, resulting in a more consistent graph.

2) K-mean clustering: 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 [27], [28]. 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.

3) Optimal number of clusters: Researchers have proposed many methods to find the optimal number of clusters for the k-mean algorithm (e.g.: Elbow method [29], Gap Statistic [30], Cross Validation [31], Silhouette method [32] and many others). In this work, the optimal number of clusters was

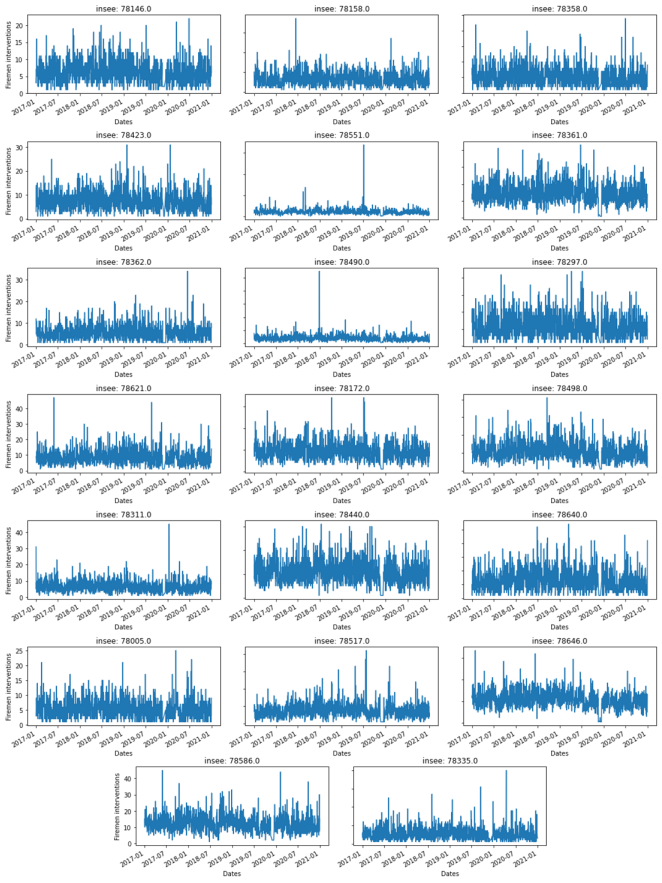

Fig. 3. Number of elements of each dictionary

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 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 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.

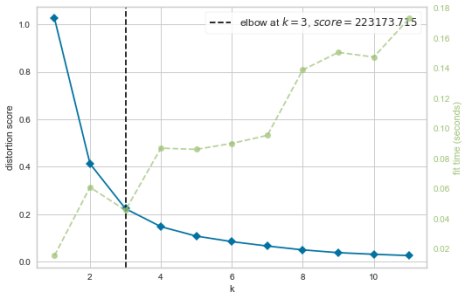

Fig. 4. Distortion score Elbow for kmeans clustering

Along with the clustering technique and the optimal number of clusters, Dynamic Time Warping (DTW) was chosen as the distance metric for Time Series [33].

4) Clustering results: The twenty insee picked up in the department of Yvelines were grouped in three distinct clusters without overlap: Cluster 0 (Figure 5), Cluster 1 (Figure 6), Cluster 3 (Figure 6). Additionally, Linear regression [34] was implemented to elucidate the mean absolute error and the mean squared error for each different cluster displayed in Table III.

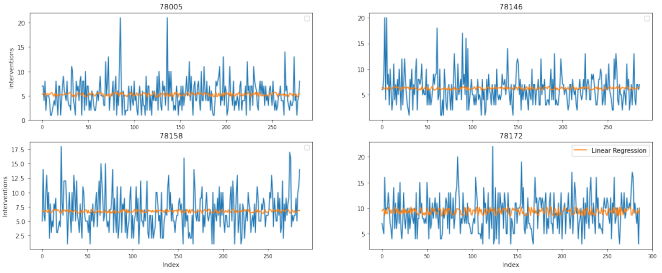

Fig. 5. Cluster0

D. 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' [35] as indicated in Table V. The reason for this step is to investigate whether the breakpoint affects the segregation of the clusters or not. More specifically,

Fig. 6. Cluster1

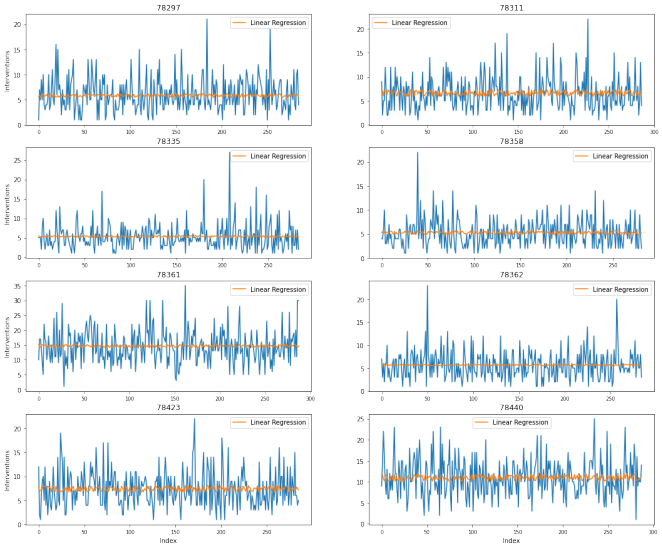

Fig. 7. Cluster2

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. This was evaluated to see if these statistics influence the cluster breakdown.

TABLE III
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

TABLE IV
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

IV. RESULTS DISCUSSION

In Section III, several assessments were made: The data were 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 5,6,7, 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

TABLE V
STATISTICS FOR DISTINCT INSEE

Cluster	Insee 78-	Population	75 or older	Farmers	Entrepreneurs	Higher intellectual professions	Employees	Retired	No professional activity
	005	21098	1051	3	292	1833	3511	2372	2249
	146	30330	2749	14	778	5965	3293	5078	3662
0	158	31306	3914	3	619	3725	891	2978	1346
	172	35656	2998	9	664	3848	4743	6276	4322
	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
1	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
	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
2	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

time series decomposition (DTS) technique is applied, which envisage 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 Figure 8.

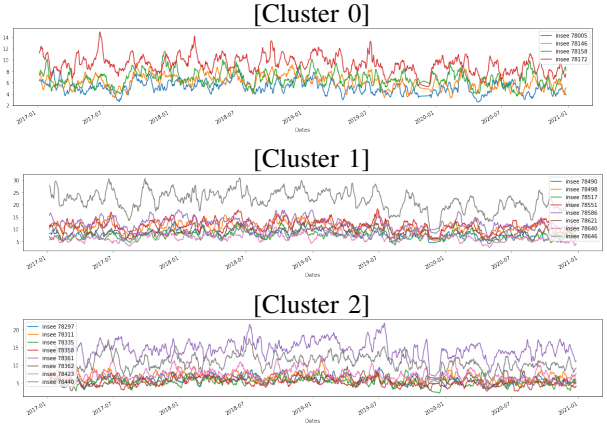

Fig. 8. Time Series Decomposition the three clusters

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.

Ultimately, 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.

V. CONCLUSION

This paper 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 number of fire interventions was discussed. Also, the outliers has been removed in 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 have 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.

In future studies, it may be possible to add population variables and geographic information for each location to the existing dataset. Also, a different technique of clustering can be applied by collecting the fire incidents in different regions of France.

VI. ACKNOWLEDGMENTS

This work has been supported by the EIPHI Graduate School (contract ANR-17-EURE-0002) and is partially funded with support from the Hubert Curien CEDRE programme n° 46543ZD.

REFERENCES

[1] D. Shaub, "Fast and accurate yearly time series forecasting with forecast combinations," *International Journal of Forecasting*, vol. 36, no. 1, pp. 116–120, 2020.

[2] N. Kulendran and M. L. King, "Forecasting international quarterly tourist flows using error-correction and time-series models," *International Journal of Forecasting*, vol. 13, no. 3, pp. 319–327, 1997.

[3] G. Nunnari and V. Nunnari, "Forecasting monthly sales retail time series: a case study," in *2017 IEEE 19th conference on business informatics (CBI)*, vol. 1. IEEE, 2017, pp. 1–6.

[4] R. Khaldi, A. El Afia, and R. Chiheb, "Forecasting of weekly patient visits to emergency department: real case study," *Procedia computer science*, vol. 148, pp. 532–541, 2019.

[5] M. Mudassir, S. Bennbaia, D. Unal, and M. Hammoudeh, "Time-series forecasting of bitcoin prices using high-dimensional features: a machine learning approach," *Neural computing and applications*, pp. 1–15, 2020.

[6] G. Zhang, D. Yang, G. Galanis, and E. Androulakis, "Solar forecasting with hourly updated numerical weather prediction," *Renewable and Sustainable Energy Reviews*, vol. 154, p. 111768, 2022.

[7] O. Raiter, "Segmentation of bank consumers for artificial intelligence marketing," *International Journal of Contemporary Financial Issues*, vol. 1, no. 1, pp. 39–54, 2021.

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

[9] S. P. Tamba, M. D. Batubara, W. Purba, M. Sihombing, V. M. M. Siregar, and J. Banjarnahor, "Book data grouping in libraries using the k-means clustering method," in *Journal of Physics: Conference Series*, vol. 1230, no. 1. IOP Publishing, 2019, p. 012074.

[10] A. Ip, S. Fong, and E. Liu, "Optimization for allocating bev recharging stations in urban areas by using hierarchical clustering," in *2010 6th International Conference on Advanced Information Management and Service (IMS)*, 2010, pp. 460–465.

[11] A. Kuhn, S. Ducasse, and T. Gîrba, "Semantic clustering: Identifying topics in source code," *Information and software technology*, vol. 49, no. 3, pp. 230–243, 2007.

[12] S. L. C. Ñahuis, C. Guyeux, H. H. Arcolezi, R. Couturier, G. Royer, and A. D. P. Lotufo, "Long short-term memory for predicting firemen interventions," in *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2019, pp. 1132–1137.

[13] J.-F. Couchot, C. Guyeux, and G. Royer, "Anonymously forecasting the number and nature of firefighting operations," in *Proceedings of the 23rd International Database Applications & Engineering Symposium*, 2019, pp. 1–8.

[14] C. Guyeux, J.-M. Nicod, C. Varnier, Z. Al Masry, N. Zerhouni, N. Omri, and G. Royer, "Firemen prediction by using neural networks: a real case study," in *Intelligent Systems Conference (IntelliSys 2019)*, London, United Kingdom, sep 2019, pp. 541 – 552. [Online]. Available: https://publiweb.femto-st.fr/tntnet/entries/15539/documents/author/data

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

[16] S. Cerna, C. Guyeux, H. H. Arcolezi, R. Couturier, and G. Royer, "A comparison of lstm and xgboost for predicting firemen interventions," in *World Conference on Information Systems and Technologies*. Springer, 2020, pp. 424–434.

[17] S. Cerna, C. Guyeux, G. Royer, C. Chevallier, and G. Plumerel, "Predicting fire brigades operational breakdowns: A real case study," *Mathematics*, vol. 8, no. 8, p. 1383, 2020.

[18] S. Cerna, H. H. Arcolezi, C. Guyeux, G. Royer-Fey, and C. Chevallier, "Machine learning-based forecasting of firemen ambulances' turnaround time in hospitals, considering the covid-19 impact," *Applied soft computing*, vol. 109, p. 107561, 2021.

[19] R. Elias Mallouhy, C. Guyeux, C. A. Jaoude, and A. 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*. Springer, 2022, pp. 3–12.

[20] R. Elias Mallouhy, C. Guyeux, C. Abou Jaoude, and A. Makhoul, "Time series forecasting for the number of firefighters interventions," in *International Conference on Advanced Information Networking and Applications*. Springer, 2021, pp. 39–50.

[21] R. E. Mallouhy, C. Guyeux, C. A. Jaoude, and A. Makhoul, "Forecasting the number of firemen interventions using exponential smoothing methods: a case study," in *International Conference on Advanced Information Networking and Applications*. Springer, 2022, pp. 579–589.

[22] R. Mallouhy, C. Guyeux, C. Abou Jaoude, and A. Makhoul, "Machine learning for predicting firefighters' interventions per type of mission," in *8th IFAC/IEEE International Conference on Control, Decision and Information Technologies (CoDIT 2022)*, Istanbul, Turkey, may 2022. [Online]. Available: https://publiweb.femto-st.fr/tntnet/entries/18971/documents/author/data

[23] R. Elias Mallouhy, C. Guyeux, C. Abou Jaoude, and A. Makhoul, "Predicting fire brigades' operations based on their type of interventions," in *18th IEEE International Wireless Communications and Mobile Computing Conference (IWCMC 2022)*, Dubrovnik, Croatia, may 2022. [Online]. Available: https://publiweb.femto-st.fr/tntnet/entries/19009/documents/author/data

[24] C. Guyeux, A. Makhoul, and J. Bahi, "How to build an optimal and operational knowledge base to predict firefighters' interventions," in *Intelligent Systems Conference (IntelliSys 2022)*, Amsterdam, Netherlands, sep 2022. [Online]. Available: https://publiweb.femto-st.fr/tntnet/entries/18941/documents/author/data

[25] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[26] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[27] D. Kaur and K. Jyoti, "Enhancement in the performance of k-means algorithm," *International Journal of Computer Science and Communication Engineering*, vol. 2, no. 1, pp. 29–32, 2013.

[28] A. Bansal, M. Sharma, and S. Goel, "Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining," *International Journal of Computer Applications*, vol. 157, no. 6, pp. 0975–8887, 2017.

[29] D. Marutho, S. Hendra Handaka, E. Wijaya, and Muljono, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," in *2018 International Seminar on Application for Technology of Information and Communication*, 2018, pp. 533–538.

[30] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[31] W. Fu and P. O. Perry, "Estimating the number of clusters using cross-validation," *Journal of Computational and Graphical Statistics*, vol. 29, no. 1, pp. 162–173, 2020.

[32] A. Lengyel and Z. Botta-Dukát, "Silhouette width using generalized mean—a flexible method for assessing clustering efficiency," *Ecology and evolution*, vol. 9, no. 23, pp. 13 231–13 243, 2019.

[33] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[34] X. Yan and X. Su, *Linear regression analysis: theory and computing*. World Scientific, 2009.

[35] C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline change point detection methods," *Signal Processing*, vol. 167, p. 107299, 2020.