INM430 - Tiny DS Project Report

Predicting response times of the Paris Fire Brigade vehicles

Célia Detrez

1 Statement Analysis Domain, Questions and Plan

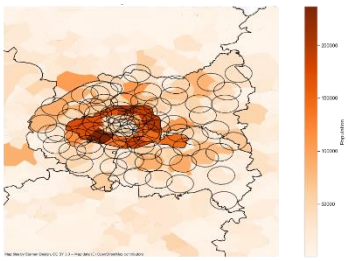
The data is extracted from the 2018 data provided by the fire brigade of Paris **[1]**. Understanding and optimizing the response times of emergency services is primordial to reduce the casualties and damages **[2]**. The longer an emergency team takes, the worst the situation could be. The real response time is often higher than the response time predicted by the Open Source Routing Machine. This observation highlights some latent features in the prediction. Identifying these factors would be a great help for enhancing the real-time decision-making process. The location of the intervention, meteorological and traffic factors - highly correlated with the peak hours - are expected to influence the quality of this prediction **[3]**. This tiny DS project tries to find patterns in the data needed to understand the observed response times. Analysing the data should provide insights on the features that affect the time between the departure and the presentation of the fire brigade and allow to identify different groups of response times.

Firstly, the main table is merged with daily weather data provided by Samson **[4]**. Missing values inspection is performed. As the original size of the data is large, the columns with more than 90% of missing data and the samples with more than 10% missing values of the features are dropped. A dataset of 61,515 samples and 53 features with 19 categorical and 34 numerical features is obtained. The numerical columns are studied closer through the analysis of their distribution and their mutual R^2 and Spearman correlations. Dependent and independent attributes are identified. Irrelevant features are detected and ignored if so. Then, the values of the categorical columns are transformed to employ them in further analysis. The final data set has 818 features. The influence of the time-related features on the response times and the number of interventions is analysed. Then, a feature dependency analysis is performed by plotting the influence of reasons, location of the event, the vehicle and the weather. Modelling is computed by deriving a feature splitting the response times into 3 groups. Geographical patterns are checked to decide if the latitude and longitude points should be included **[5,6,7,8]**. Then, different dimension reduction techniques as PCA, LDA, MCA, and T-SNE are considered to deal with many features and help identify groups. A K-means algorithm is used to cluster the data based on the groups and the independent features. The quality of each reduction technique is assessed by fitting a random forest classifier on a train set (80%) and evaluating it on a test set (20%). Finally, the characteristics are investigated by plotting the influence of the most significant features of the random forest on the response times for each group. A linear regression model for each cluster on the response times is fitted to put under the spotlight the most significant features of each group of response times. Those features should coincide with the most important ones for the random forest. Results can be observed on a map.

2 Findings and Reflections

2.1 Geographical patterns

Response times are linked to the distance that the emergency vehicles must drive to reach the location of the intervention. With a 2km interventions radius, the intervention areas of the fire brigades overlap when the density of population is high. Hence, emergency services are supposed to arrive within about the same time. Consequently, the position-related features will not be included in the model.



a. Population distribution

b. Fire brigade interventions at midnight on Fridays

Fig. 1. Fire brigade interventions areas

2.2 Feature reduction

As the data set has many features, namely 818, reduction techniques help identify the clusters. 4 different reduction techniques have been implemented as Linear Discriminant Analysis. Principal components analysis and Factor Analysis for Mixed Data methods are highly sensitive to outliers and

have not allowed visualising the different groups **[11]**. T-SNE reduction method works on a large dataset but has a lower efficiency when the number of features is too high **[9]**. However, even if the data does not respect the multivariate normality and the homoscedasticity assumptions, Linear Discriminant Analysis reduction, as plotted below, has been successful. The Multiple Correspondence Analysis method has been quite effective too.

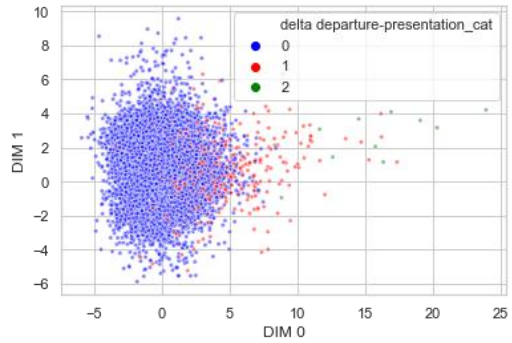


Fig. 2. LDA dimensions

2.3 Clustering

K-means method is implemented to cluster the samples according to their delta departure-presentation group and the independent features. K-means using LDA features instead of the numerical features clustering is compared to K-means with all the independent features.

To choose between the two k-means clustering methods, a random forest classifier is fitted. K-means with LDA features clustering method has the highest accuracy. It can be explained by the class imbalance produced by the MCA and no reduction K-means. Random Forest is based on decision trees that are highly sensitive to class imbalance.

	0	1	2
No reduction	58735	198	2582
LDA	2242	1559	57714
MCA	58735	2582	198
LDA MCA	9445	47785	4285

Table 1. K-means clusters size

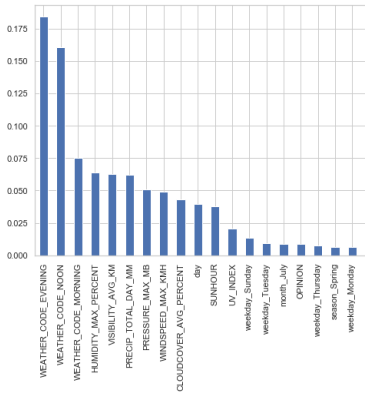


Fig. 3. Random Forest features importance

The features importance bar plot of the random forest implemented to predict the K-means clusters with LDA highlights that the WEATHER and the LDA features are significant in the model.

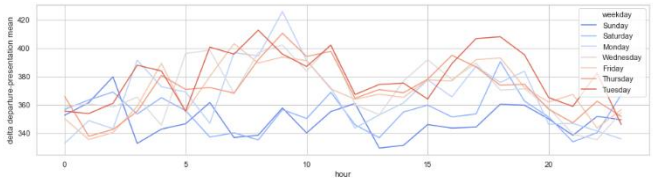
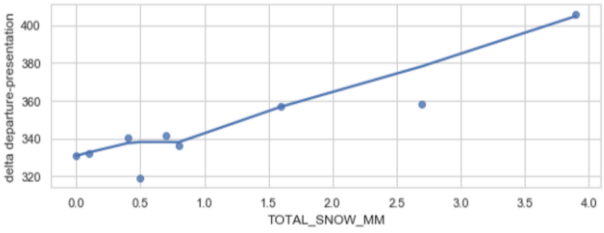
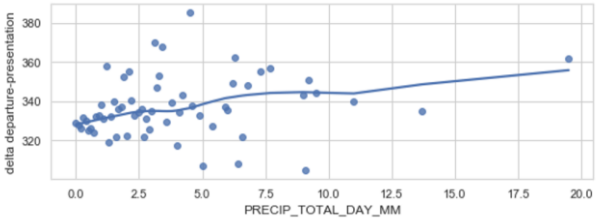


Fig. 4. delta departure-presentation mean per hour by weekday

During the working days and peak hours, the delta departure-presentation is higher than during the weekend. It could be due to the high traffic during these hours.



a. Total snow



b. Total precipitations

Fig. 6. Weather influence on delta departure-presentation

Emergency vehicles take more time during “bad” weather day than during “good” weather day. It can be explained by higher demand on these days and by unusual circumstances.

2.4 Modelling the response times for each cluster

The linear regression highlights the different features statistically significant for each cluster. The graphs below show some of the features allowing to define different types of response times.

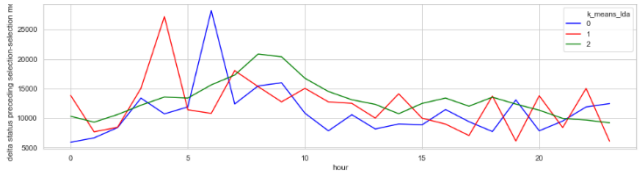


Fig. 7. delta status preceding selection-selection mean per hour by K-means LDA clusters

The delta status preceding selection-selection mean in group 2 is higher on average than the ones in group 1 and 0. It can be made the hypothesis that groups 1 and 0 are related to urgent events. In the early morning the times of the selections

for groups 1 and 0 increase, it can be supposed that the firemen on duty are not as much as during the day and for some intervention they need specific equipment.

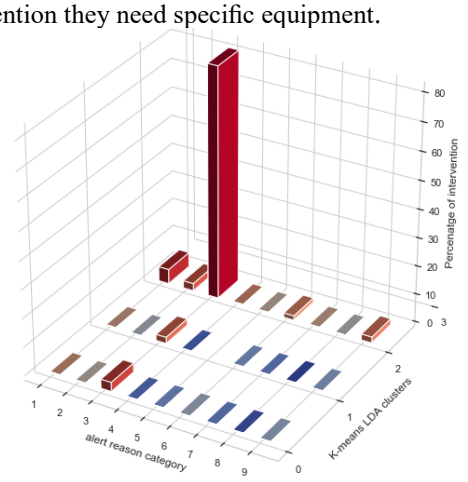


Fig. 8. Percentage of interventions per alert reason category and K-means LDA clusters

Most of the interventions are for an alert reason category 3, 1, or 2. However, alert reason category 5 is not observed for group 1 interventions. The response times in group 2 are higher on average than the ones in groups 0 and 1.

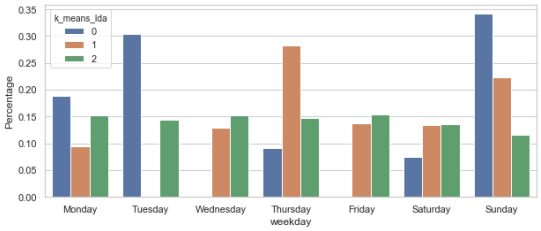


Fig. 9. Percentage of intervention per weekday by K-means LDA clusters

Most of the intervention of group 1 are on Thursday and Sunday whereas group 0 interventions are mostly on Tuesday, Sunday, and Monday. The interventions on Sunday could be linked to the fact that a lot of people go out during the weekend. As it could also explain the Thursday interventions of group 1. The interventions of group 0 at the beginning of the week could be linked to the lack of attention at the working place. The U.S. Bureau of Labour Statistics stated that Monday is the most dangerous day for workplace accidents **[12]**.

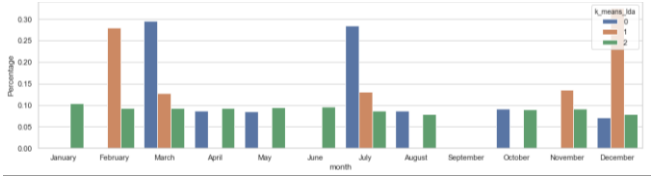


Fig. 10. Percentage of intervention per month by K-means LDA clusters

During "cold" months, the number of interventions of group 1 is higher than during the other months. During July and March, the number of interventions of group 0 is higher. It matches with the change of seasons. Group 0 interventions happen only during "cold" months and "hot" months. The number of interventions of group 0 is the same all over the year.

2.5 Conclusion and future work

This project has highlighted three groups of response times that can help the Fire Brigade of Paris:
- the first one, group 0, stands for the week-ends and beginning of the week unusual interventions.
- the second one, group 1, symbolises the working unusual interventions.
- the third one, group 2, represents everyday life interventions. It is the usual workflow across the year.

The two first groups are observed more frequently during the cold and the warm spells.

However, this analysis has some limitations. As it has been stated previously, the data does not verify all the assumptions for LDA, and an alternative method should be used. Moreover, there is no samples of September interventions. It could have influenced the results. In addition, as the journey is not provided for each data point, the traffic has not been included but as the hour of day influences the response times it could be very interesting to include at least a traffic index of each day.

To further this analysis, it could be enriching to include the numbers of people on duty in the model as it could affect significantly the response times.

Table of word counts

Statement Analysis, Problem and Plan	495
Findings and Reflections	987

References

[1] Paris Fire Brigade, "Challenge," Github.io, 2019. [Online]. Available: https://paris-fire-brigade.github.io/data-challenge/challenge.html. [Accessed: 22-Oct-2019]

[2] M. Bandyopadhyay and V. Singh, "Development of agent-based model for predicting emergency response time," Perspectives in Science, vol. 8, pp. 138–141, Sep. 2016.

[3] J. R. Blum, A. Eichhorn, S. Smith, M. Sterle-Contala, and J. R. Cooperstock. "Real-time emergency response: improved management of real-time information during crisis situations", Journal on Multimodal User Interfaces, vol. 8, no. 2, pp. 161–173, Dec. 2013.

[4] T. Samson, "Météo à Paris", Historique Météo, 2019. [Online]. Available: https://www.historique-meteo.net/france/ile-de-france/paris/. [Accessed: 02- Dec- 2019].

[5] "Liste des casernes à Paris et dans les départements de la petite couronne", *Data.gouv.fr*, 2019. [Online]. Available: https://www.data.gouv.fr/fr/datasets/liste-des-casernes-a-paris-et-dans-les-departements-de-la-petite-couronne-551678/#_. [Accessed: 03- Dec- 2019].

[6] "Les communes d'Île-de-France", *Data.gouv.fr*, 2019. [Online]. Available: https://www.data.gouv.fr/fr/datasets/les-communes-d-ile-de-france-idf/. [Accessed: 03- Dec- 2019].

[7] "Arrondissements", *Data.gouv.fr*, 2019. [Online]. Available: https://www.data.gouv.fr/fr/datasets/arrondissements-1/. [Accessed: 03- Dec- 2019].

[8] "Données communales sur la population d'Île-de-France", *Data.gouv.fr*, 2019. [Online]. Available: https://www.data.gouv.fr/fr/datasets/donnees-communales-sur-la-population-d-ile-de-france-idf/#_. [Accessed: 03- Dec- 2019].
[9] M. Wattenberg, F. Viégas, and I. Johnson, "How to Use t-SNE Effectively," Distill, vol. 1, no. 10, Oct. 2016.
[10] "Les engins", Pompiersparis.fr, 2019. [Online]. Available: https://www.pompiersparis.fr/fr/operationnelle/engin. [Accessed: 30- Oct- 2019].
[11] Ş. Büyüköztürk and Ö. Çokluk-Bökeoğlu, "Discriminant Function Analysis: Concept and Application," Eurasian Journal of Educational Research, vol. 33, pp. 73–92, 2008.
[12] "Nonfatal Occupational Injuries and Illnesses Requiring Days Away From Work," U.S. Bureau of Labor Statistics, 2016.