Université Paris Descartes
Faculté d'informatique

Rapport de projet

Module : Apprentissage non-supervisé

Master 2 : MLDS

Étude et analyse temporelle d'un jeu de données des taux d'occupations de parkings

- Réalisé par :

  Vanna Boungnalith

  Nazim Messous

  Wissam Benhaddad

7-11-2019

1 Préambule

Dans ce projet, nous allons analyser un dataset représentant la disponibilité de plusieurs parkings à Birmingham. Il y a en tout 30 parkings où le nombre de places de parking est censé être mis à jour toutes les 30 minutes de 08h00 à 16h30 pendant 77 jours, du 04 Octobre 2016 au 19 Décembre 2016.

Pour cela, l'étude va être divisée en quatre parties. Nous allons d'abord préparer les données afin que celles-ci soient exploitables et ne posent pas de problèmes lors de leur analyse. Nous ferons ensuite une première étude descriptives des données nettoyées avant de nous concentrer sur l'analyse du comportement hebdomadaire des parkings. Enfin, nous essayerons de regrouper les parkings en classe homogènes à partir des résultats que nous avons accumulé au long de ce projet.

2 Préparation des données

Dans cette section nous allons nous attarder sur l'aspect organisationnel des données.

2.1 Première impression sur le jeu de données

Après plusieurs observations du dataset original, plusieurs points ont été retenus et exige de notre part une modification du dataset d'origine:

- **Absence de mesures:** Chaque parking doit contenir 18 mesures pendant 77 jours soit 1386 mesures par parking sur la période de temps donnée. Or, aucun parking n'a atteint ce nombre de mesure. Le parking avec le plus de mesures ne possédait que 1312 mesures.
- **Date des mesures:** La plupart des mesures ne sont pas prises exactement toutes les demi-heures.
  Exemple: Une mesure à été prise à 8h59min33sec et non 9h00min00sec
- **Mesures aberrantes:** Certaines mesures montre que le nombre de véhicules stationnés à un instant d'un parking t dépasse la capacité maximal du parking concerné. D'autres mesures sont parfois négatives, ce qui est impossible.
- **Parkings avec trop peu de données:** Deux parkings possèdent très peu de données comparé aux autres parkings (88 et 162 valeurs).
- **Précense de doublons:** Certaines mesures sont exactement identiques, nous avons pu compter 53 doublons.

Il est important de résoudre les problèmes mentionnés car ils pourraient rendre l'analyse du dataset très difficile. Les problèmes observés pourraient être du à des capteurs non fiable ou défectueux.

2.2 Data cleaning: Étapes de préparation du dataset

Nous sommes passé par trois étapes de modifications du dataset afin de bien préparer les données:

- **Première étape:** On supprime les deux parkings avec trop peu de mesures. Il serait trop compliqué des les introduire dans notre étude car il manque pour ces deux parkings près de 90% de mesures. On remarque également que toutes les valeurs négatives appartenaient à un des parkings supprimés.
  Toutes les valeurs dépassant la capacité maximum d'un parking sont remplacés par la valeur de la capacité maximum.

- **Deuxième étape:** On réorganise les horaires de prise de mesure afin de faciliter le partionnement des données. Chaque instant d'une prise de mesure est arrondi à la demi-heure la plus proche.

SystemCodeNumber	Capacity	Occupancy	LastUpdated
BHMBCCMKT01	577	61	2016-10-04 07:59:42
BHMBCCMKT01	577	64	2016-10-04 08:25:42
BHMBCCMKT01	577	80	2016-10-04 08:59:42
BHMBCCMKT01	577	107	2016-10-04 09:32:46

SystemCodeNumber	Capacity	Occupency	LastUdapted
BHMBCCMKT01	577	61	2016-10-04 08:00:00
BHMBCCMKT01	577	64	2016-10-04 08:30:00
BHMBCCMKT01	577	80	2016-10-04 09:00:00
BHMBCCMKT01	577	107	2016-10-04 09:30:00

Figure 1: Changement sur la colonne "Lastly update"

  On supprime également tous les doublons.

- **Troisième étape:** Tous les jours, chaque parking doit contenir 1386 mesures. Nous avons donc effectué un remplissage des valeurs manquantes: on remplace chaque valeur manquante par la moyenne locale. $(\frac{valeurprec+valeursuiv}{2})$
  Si une journé à plus de 5 mesures manquantes on remplace ses valeurs par celles de la semaine suivante.

Le data cleaning est absolument primordial si on veut travailler sur ces données. Cela nous permet d'avoir un dataset exploitable lors de l'utilisation de méthodes de description statistique et des méthodes d'apprentissage non supervisé.

3 Statistiques descriptives

Résumé des parkings

Les 28 parkings que nous avons gardé ont tous des tailles différentes. Certains sont assez petit avec une capacité n'excédant pas 500 places. D'autres ont une capacité beaucoup plus élevé allant jusqu'à 4675 places. Certains parkings sont à côté de lieux très fréquentés tels que des centres commerciaux ou des grandes gares ce qui explique une grande capacité.

```
> summary(liste_capacity)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  220.0   557.8   856.0  1392.1  1942.2  4675.0
```

Figure 2: Résumé des capacités des parkings étudiés

Taux d'occupation moyen des parkings

Afin d'avoir une meilleure visualisation de la fréquentation des parkings, il peut être intéressant, voire même nécessaire de calculer le taux de fréquentation des parkings en divisant la capacité à un instant t par la capacité maximale.

```
> summary(liste_moy_per)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1650  0.3398  0.5290  0.4916  0.6208  0.7426
> liste_moy_per
 [1] 0.2852041 0.4337357 0.6652855 0.7425622 0.6191669 0.4838091 0.3214366 0.6378158 0.6009183 0.6945038 0.4654582 0.7043500
[13] 0.6681573 0.5940673 0.1964951 0.6028800 0.6258862 0.4518747 0.1649682 0.2522918 0.5685049 0.1938487 0.4727700 0.5881571
[25] 0.3436009 0.5031824 0.3284214 0.5549096
```

Figure 3: Résumé des taux d'occupations

Sur la figure ci-dessus, nous avons calculé le taux d'occupation moyen de chaque parkings sur l'intervalle de temps d'étude, c'est-à-dire 77 jours. Il a fallut calculer le taux d'occupation de chaque parking à chaque mesure et d'en tirer la moyenne.

Taux d'occupation moyen de parkings par semaine

Nous avons également travailler sur les taux d'occupations de parking par semaine afin de pouvoir visualiser les différences entre les parking.
Sur la figure suivante, nous comparons deux parkings avec des comportement différents: le quatrième parking qui a un taux d'occupation assez variable car celui-ci monte fortement lors de la 5eme semaine et le dixième parking qui est normalement fréquenté, avec des variations moins forte.

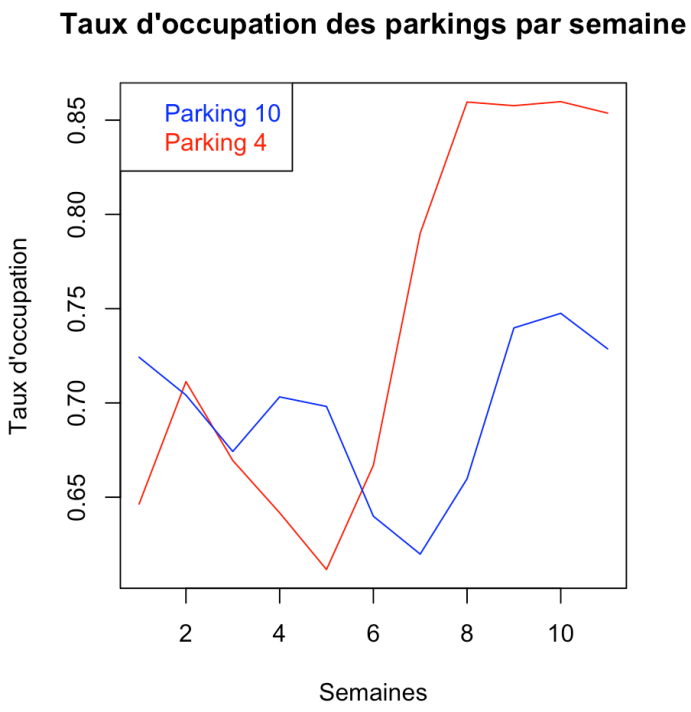

Figure 4: Changement sur la colonne "Lastly update"

4 Analyse du comportement hebdomadaire

Dans cette section nous traiterons de l'analyse hebdomadaire des séries temporelles.

Chaque ligne représente une semaine pour un parking, les 11 premières lignes représentent les 11 semaines du premier parking et ainsi de suite.

4.1 Kmeans

4.1.1 distance

Utilisation d'une distance euclidienne

$$d(x, y) = \sum_{i=1}^{p} (x_i - y_i)^2$$

4.1.2 Choix du nombre de cluster

Pour le choix du nombre de cluster on a utilisé le système de vote de la librairie Nbclust, qui utilise au total 30 critères pour décider du nombre de cluster optimaux avec comme critères d'exemples: "gap", "silhouette" ou encore "gamma".
-5 critère ont voté pour 2 cluster
-12 critère ont voté pour 3 cluster
-4 critère ont voté pour 8 cluster
On choisit donc 3 cluster.

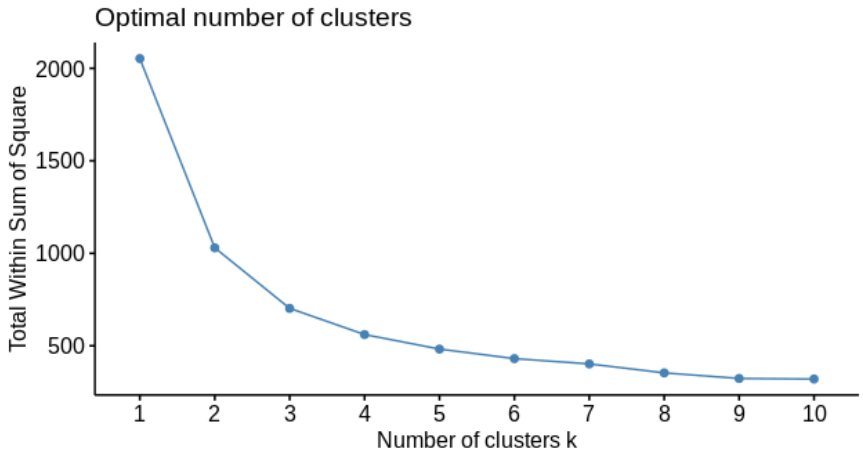

Figure 5: Choix du nombre de clusrer par la méthode du coude

4.1.3 Visualisation des semaines associées aux différents parking sur les clusters :

Les parkings 2, 4, 10, 11, 14, 20 et 23 ont des mesures (semaines) dans plus d'un cluster Le reste des parkings ont leur 11 semaines dans un seul cluster Cette information est importante car elle prouve que les parkings ont globalement un comportement similaire sur l'ensemble des semaines. Les 7 parkings ayant des mesures sur différents clusters ont leurs mesures sur pas plus de 2 clusters à l'exception du parking 11.

Interpretation du premier cluster :

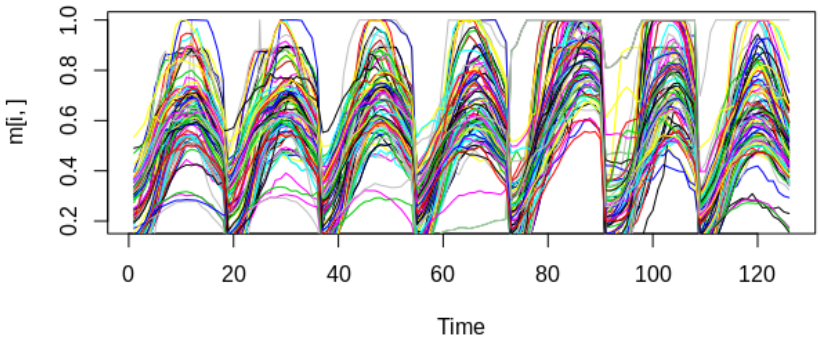

Figure 6: Cluster 1 obtenue avec K-means

Il y a 96 semaines dans ce cluster.

- Taux d'occupancy le week end: 0.5859115
- Taux d'occupancy les jours de travail: 0.5125178
- Taux d'occupency sur toute la semaine de 0.5334874

On voit clairement que les semaines de ce cluster ont un taux d'occupant de 50%, avec un léger accroissement le weekend. Il s'agit là de semaines de parkings moyennement fréquentés.
Ce léger accroissement peut être expliqué par le fait que ces parkings sont localisés près de centres commerciaux ou de loisirs ou d'autres endroits à forte fréquentation durant le weekend.
Nous avons pour ce cluster des parkings comme Bhmbccmkt01(markets) ou BHMMBMMBX01(mail box), tous les deux des centres commerciaux à Birmingham.

Interpretation du deuxième cluster:

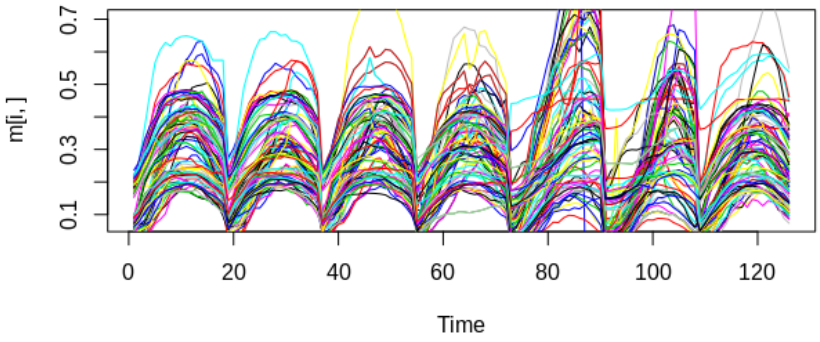

Figure 7: Cluster 2 obtenue avec K-means

Il y a 89 semaines dans ce cluster.

- Taux d'occupancy le week end: 0.2566391
- Taux d'occupancy les jours de travail: 0.264372
- Taux d'occupency sur toute la semaine de 0.2621626

Parkings à faible fréquentation, généralement on retrouve des semaines de parkings situés près d'habitations sans bureau ou des lieux qui avec peu d'activité la journée.
Par exemple. Le parking BHMBCCPDC01(prôche du Paradise Circus, une zone culturel) à ses 11 semaines dans ce cluster car les gens se rendent au Paradise Circus le soir et non la journée. De même pour le parking BHMBCCTHL01 qui est un parking pour la salle de concert du Hall Town et qui est donc fréquenté le soir.

Interpretation du troisième cluster :

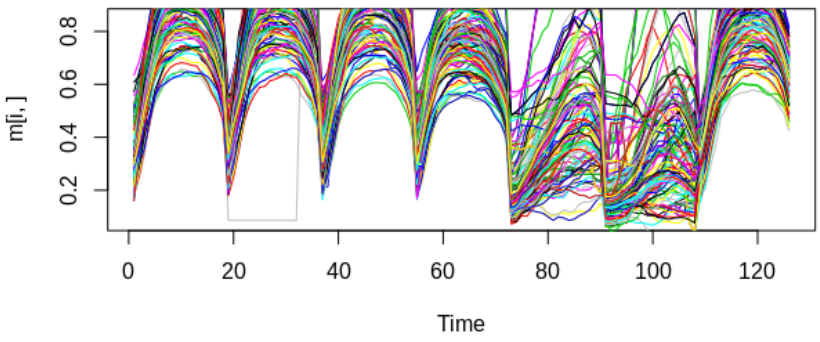

Figure 8: Cluster 3 obtenue avec K-means

Il y a 123 semaines dans ce cluster

- Taux d'occupancy le week end: 0.3448578
- Taux d'occupancy les jours de travail: 0.7365292
- Taux d'occupency sur toute la semaine de 0.6246231

Les semaines de ce cluster ont le taux d'occupancy le plus élevé des trois clusters, 62% en moyenne, avec une nette augmentation pour les jours de travail et une baisse drastique le weekend. Cette différence peut être expliquée par le fait que ces semaines appartiennent à des parkings se trouvant à proximité de bureaux ou de lieux de travail de façon générale. Cela explique pourquoi ces parkings sont très peu fréquentés le week-end. Nous pouvant citer le parking BHMEURBRD01(Broad Street, Birmingham) le quartier d'affaire de Birmingham ou il y a plusieurs grande banques notamment et divers bureaux.

Remarque importante La méthode de K-means avec une distance DTW ainsi que l'utilisation d'un k-médoids avec une initialisation à l'aide CAH (critère de Ward) a donné un nombre de clusters identique avec des clusters semblables à ceux produit par notre k-means avec distance euclidienne.
Par exemple, les cinquième et septième parking n'ont pas leurs 11 semaines dans un seul cluster avec la méthode k-means distance DTW. Alors que le deuxième parking a ses 11 semaines dans le premier cluster. Autre remarque importante: certains parkings ont des semaines dans d'autres clusters notamment vers la fin de l'année vers décembre, et vers le premier novembre cela étant dû à l'approche des fêtes de fin d'année.

4.2 Cartes Auto-Organisatrices

4.2.1 Paramètres

- Grille de sortie de taille 10*10
- La grille de sortie est topologie hexagonale
- Fonction de voisinage du type Gaussienne

4.2.2 Interprétation des résultats

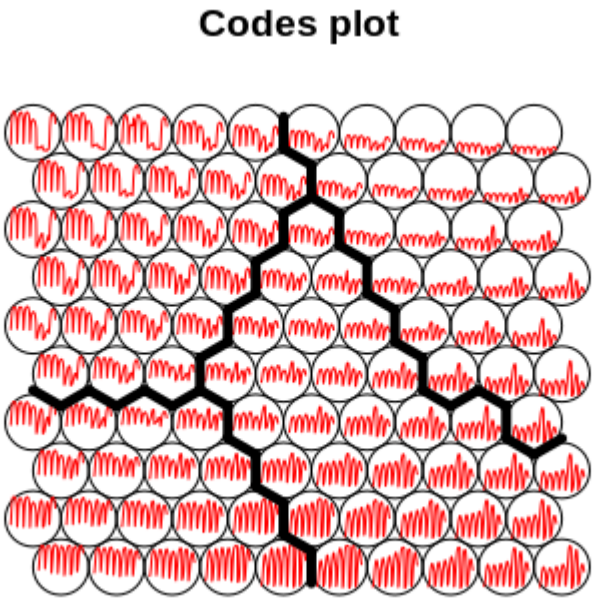

Figure 9: Classification des séries temporelles hebdomadaire selon l'algorithme SOM

On se basant sur le plot des codes et le pattern dans la distribution des échantillons et des variables on peut distingué 4 classes.

En bas à droite Le cluster connaît un taux d'occupancy assez conséquent avec un important pic le weekend et surtout le samedi

En haut à droite Cluster contenant les semaines très faibles taux d'occupancy mais stable, avec quelqueélements mal classé qui aurait dû être dans le cluster d'a coté.

En haut à gauche Cluster avec un assez fort taux d'occupancy mais surtout caractérisé par une baisse de ce taux durant le week end.

En bas à gauche Cluster contenant les semaines à fort taux d'occupancy et stable sur la semaine.

5 Analyse comportemental des parkings

Durant cette section, nous allons nous intéresser à la classification des différents parkings selon leur comportement respectif , c.à.d le taux d'occupation sur la durée totale des mesures.

5.1 Préparation des données

Avant de commencer l'analyse, il est tout d'abord nécessaire de transformer la matrice des données d'origine en une version plus compacte notée $\mathcal{M}_p$. Pour cela nous allons regrouper chaque mesure d'un même parking dans une ligne dédiée. Les dimensions de la nouvelle matrice seront donc $n \times p$ où :

- n : Le nombre de parkings, en l'occurrence 28
- p : Le nombre de mesures sur 11 semaine, soit $18 * 11 * 7 = 1386$.

Nous pouvons aussi réduire encore plus la fréquence des séries temporelles en prenant la moyenne quotidienne des 18 mesures d'un seul jour, ce qui donne une matrice $\mathcal{M}_p$ de taille 28×77

Chaque ligne est en fait une série temporelle qui décrit un parking durant la durée des mesures.

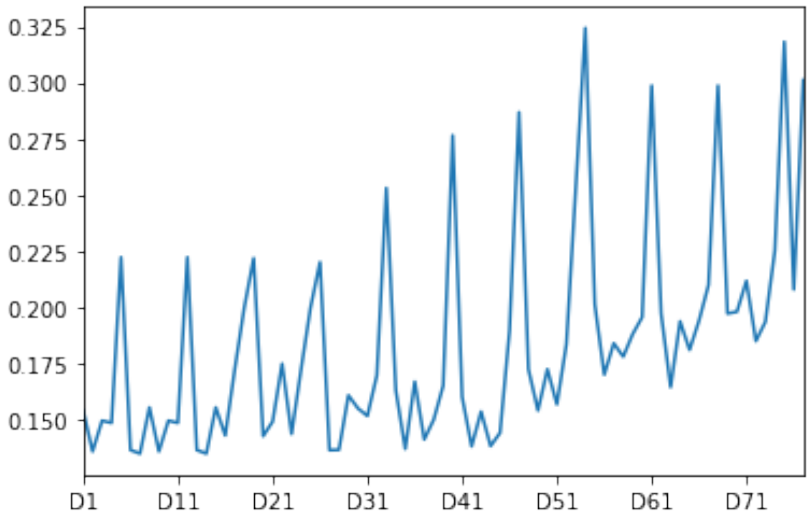

Figure 10: Variation quotidienne du taux d'occupation du parking 7

5.2 Classification avec l'algorithme K-means

Pour une première méthode, nous avons décidé d'utiliser l'algorithme **K-means** avec le critère d'inertie à optimiser. La distance utilisée est la distance Dynamic Time Warping (DTW). Nous nous sommes porté sur ce choix de méthode pour une raison expérimentale. Les données qui sont de nature numérique n'ont donc pas posé de problèmes pour le calcul des représentants.

Pour le choix du nombre de classes, nous avons tout d'abord opté pour la méthode du coude. Après analyse du graphe présent dans la figure 15. Le nombre optimal de clusters a choisir est soit égal à 3 soit égal à 2.

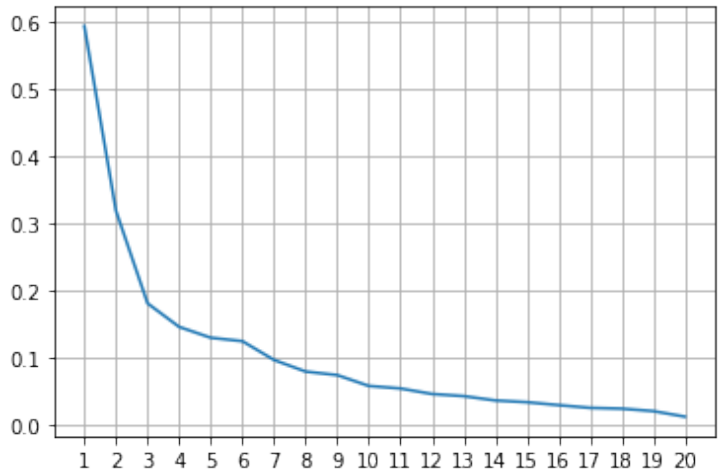

Figure 11: Variation de l'inertie totale selon le nombre de clusters K choisi

Nous avons ensuite utilisé la librairie **NbClust** précédemment citée pour consolider ce choix. parmi les 20 critères applicables, 6 ont voté pour $K = 10$, 10 ont voté pour $K = 2$, 3 vote pour $K = 3$ et un seul pour $K = 4$. Donc par vote de majorité, 2 semble, d'après les deux méthodes, le choix le plus judicieux.

Après avoir déterminer le nombre de centres à initialiser, nous avons pu lancer l'algorithme TemporalKmeans issue du package **tslearn** de Python. Les paramètres a donner sont une matrice de séries temporelles, en l'occurence notre matrice $\mathcal{M}_p$, le nombre de clusters $K = 3$ et la métrique à utiliser qui est donc la distance DTW.

Plusieurs interprétations sont possibles selon qu'on s'intéresse à un comportement hebdomadaire, c.à.d la façon dont le taux d'occupation varie selon les semaines et donc selon les tranches de périodes. Ou selon les jours de la semaine, pour analyser le comportement durant les weekends ou en fin de semaine.

Nous allons commencer par une analyse par jour. Après avoir lancé l'algorithme, les résultats sont comme suit :

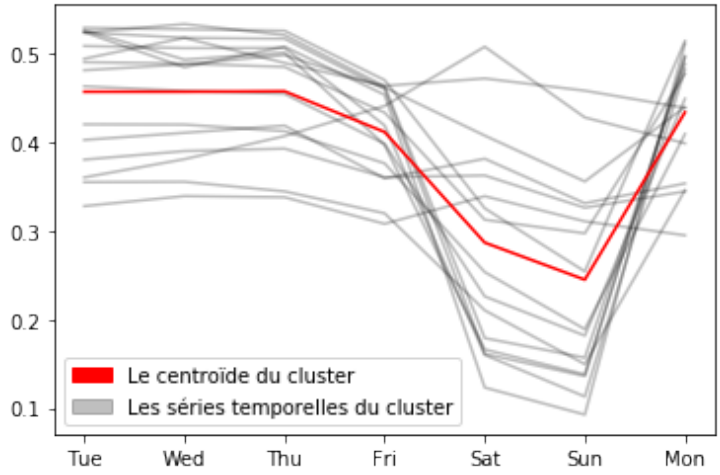

Figure 12: Séries temporelles du cluster 1

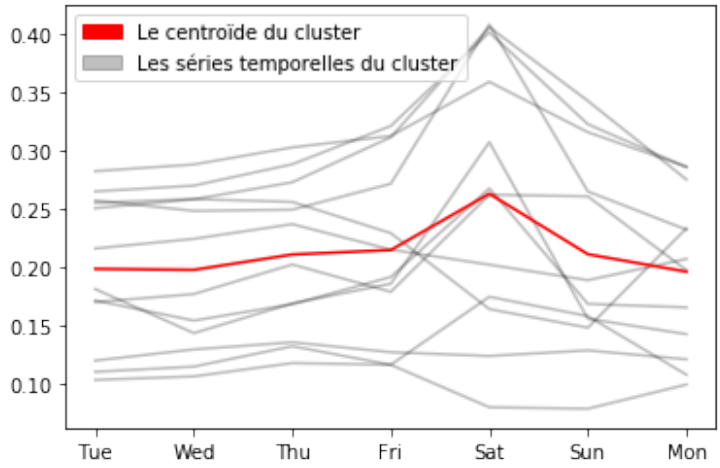

Figure 13: Séries temporelles du cluster 2

Une analyse possible pour ces deux clusters serait que le premier regroupe les parkings qui ont tendance à se vider en fin de semaine. Indiquant probablement une proximité à un quartier d'affaire où beaucoup de sièges/bureaux de sociétés se trouvent. Ces parkings auront donc tendance à être remplis toute la journée durant les jours de semaine (donc les jours de travail). Cependant cette interprétation reste basée sur une réduction par une moyenne sur les journées et les mesures des journées. Ainsi, quelques parkings attachés à des centres commerciaux ou centres de loisirs auront tendance à avoir un taux d'occupation

qui décroît les weekends sur l'ensemble de la journées, mais fortement concentré sur les heures matinales et/ou de l'après-midi. Cela s'explique par le fait que les clients ne laissent pas leurs voitures toute la journée au parking pour faire des courses. Contrairement au jour de semaine où les voitures sont stationnées durant la plus part des heures de travail.

De manière analogue, les parkings de l'autre cluster semblent suivre une variation opposée. Le taux d'occupation restent assez stable durant la semaine avec une légère augmentation durant le Samedi. Cela peut être expliqué par le fait que ces parkings peuvent être situés sur des zones proches des centres de loisirs (Centres commerciaux, parcs ou stades).

Quand on passe à une analyse plus large et plus détaillée, dans le sens où toutes les mesures ont été gardées, on remarque un comportement plus régulier.

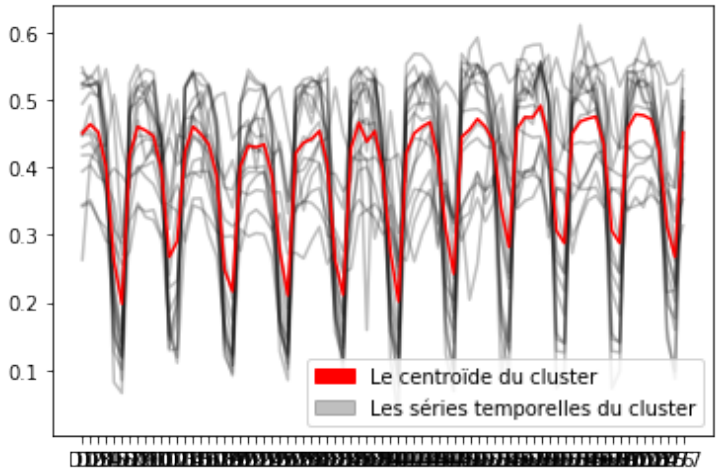

Figure 14: Séries temporelles du cluster 1

Pour cette première famille de parking, on peut remarquer une certaine régularité durant la période d'octobre à mi-décembre. On peut donc emmètre la même hypothèse que celle de l'analyse quotidienne. Et que que donc ces parkings restent assez fréquentés sur l'ensemble de la période d'analyse. Ce qui est logique car c'est la période qui suit la rentrée scolaire et qui est assez loin des période des fêtes.

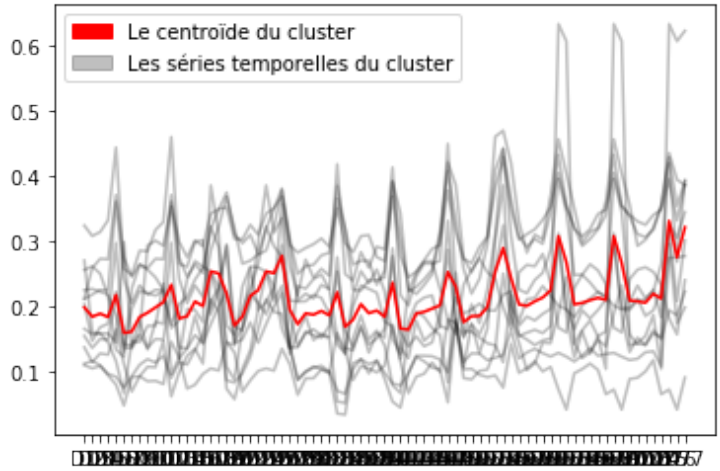

Figure 15: Séries temporelles du cluster 2

Par contre pour ce cluster, on remarque que plus on se rapproche des périodes des fêtes de fin d'années ainsi que d'Halloween, le taux d'occupation durant le weekend tend à augmenter. Cela peut se traduire par le fait que de plus en plus de personnes se ruent vers les centres commerciaux et autres zones de commerce pour faire des achats préparatifs (cadeaux, décorations, costumes, collations ...).

5.3 Classification avec l'algorithme K-means

Comme deuxième méthode, nous avons opté pour l'algorithme Self Orginzing Maps (SOM). Qui a une forte capacité de visualisation et donc offre une facilité d'interprétation non négligeable. Nous avons préalablement définit les différentes paramètres de l'algorithme SOM dans la section 4.2.1. Le choix du nombre de clusters, que nous avons choisit comme égal à 3, s'est porté sur l'analyse de différents critères. Nous passons maintenant directement à la partie expérimentale dont le résultat est explicité dans la figure 16.

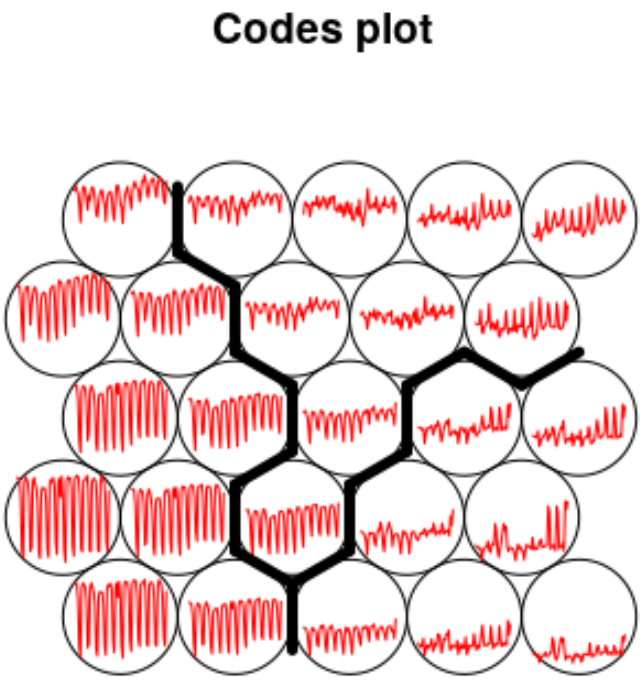

Figure 16: Classification des séries temporelles selon l'algorithme SOM

Interprétation : On peut remarquer que la majeure différence entre les séries temporelles réside dans le taux de variation du taux d'occupation. Le premier cluster, représenté sur le coté gauche de la figure 16 se compose des parkings qui se vident assez drastiquement les weekends. Cela s'accorde avec le raisonnement émis dans la section précédente.

Le deuxième cluster, qui se trouve dans la partie supérieure de la figure, regroupe lui les parkings dont le taux de variations sont assez faibles. ce cluster a pu être fusionné avec le premier dans le cas d'un analyse avec la méthode K-means qui ne prend pas en compte les taux de variation des variables.

Le troisième et dernier cluster quand à lui représente les parkings qui ont tendance à se remplir près des période de fêtes.

References