Université Paris Descartes

MLSD 2 2018-2019

Apprentissage non supervisé

Professeur :
Allou Samé
IFSTTAR

Auteurs :
Joseph Gesnouin
Yannis Tannier

Contents

1 Introduction

L'objectif de ce projet est d'exploiter la classification automatique afin de regrouper les taux d'occupation des parkings en classes homogènes. Une telle classification pourra par exemple servir à améliorer les prévisions sur ce taux d'occupation. En réalisant celles-ci classe par classe, des informations plus ciblées sur ces parkings pourraient être intéréssantes pour les directeurs de parkings, afin de mieux gérer leur parc, mais également pour les citoyens utilisant leur voiture.

Lors de ce projet, nous utiliserons certaines des techniques de classification automatique visant à classifier nos parking selon l'évolution de leur taux d'occupation de manière journalière et hebdomadaire. D'autres méthodes de classification de séries temporelles évoquées mais non vues en cours comme les modèles de mélanges seront également implémentés en annexe, en tant que complément du travail réalisé.

2 Prétraitement et exploration des données

Après avoir mis en forme les données sous forme d'un tableau individu variables, nous avons ajouté une nouvelle colonne correspondant au taux d'occupation de chacun des parkings à un instant t selon la formule: Occupancy / Capacity * 100.

De ces 30 parkings, nous en avons retiré deux que nous considerons comme outliers pendant la suite de ce rapport: le peu d'information à leur propos ne nous permettant pas de proposer une analyse porteuse d'information à leur sujet: ceux-ci disposant d'un nombre très faible de relevés comparés aux autres.

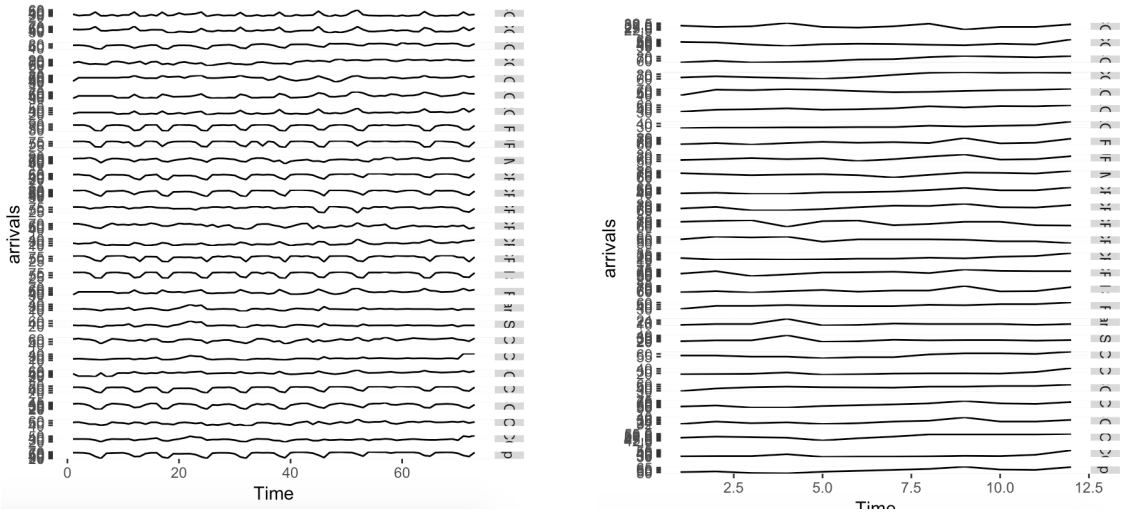

Figure 1: Représentation des 28 parkings en fonction de leur représentation journalière et hebdomadaire.

Les dates de relevés d'occupation de ces parkings se situent entre le début octobre de l'année 2016 et mi décembre 2016, pour chaque parking, nous disposons d'environ 700 relevés espacés d'une trentaine de minutes entre ces deux dates.

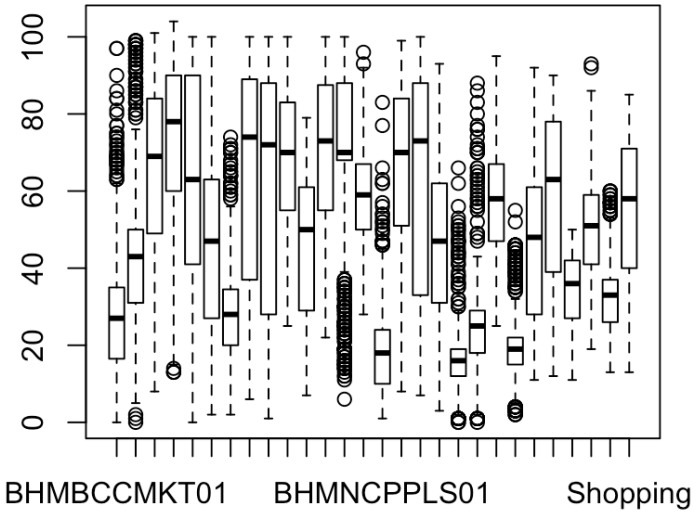

Figure 2: Boxplot de la distribution du taux d'occupation des parkings

Certains de ces parkings disposent de valeur abérantes comme des taux d'occupation superieurs à 100% pour certaines periodes données que nous avons du gérer. Ces valeurs sont probablement dues au fait d'une manipulation humaine éronée ou d'une mauvaise mise à jour de la capacité des parkings au fil du temps.

Nous remarquons des à présent une certaine disparité du taux d'utilisation des parkings en fonction des 28 parkings conservés: certains sont généralement plus occupés que d'autres par exemple.

Afin de comprendre l'évolution du taux d'occupation de ces parkings en fonction du temps, nous avons généré deux jeux de séries temporelles: grâce à la moyenne de leur occupation journalière et hebdomadaire afin de les regrouper en classes homogènes et de trouver des similarités entre ces parkings.

3 Clustering

3.1 Evolution journalière

Nous nous sommes dans un premier temps intéréssés à la classification des taux d'occupations quotidiens de la ville de Birmingham afin de repérer des similitudes au niveau de leur saisonalité.

3.1.1 Basé sur la corrélation

Dans un premier temps nous avons calculé la matrice de similarité des séries grâce aux corrélations: BHMBRCBRG02, BHMBCCMKT01, BHMBCCPST01, BHMNCPNHS01 et BHMBRCBRG01 semblent être les parkings les plus uniques en terme d'évolution du taux d'occupation: ceux-ci étant différents de la majorité des autres parkings à tout moment. La distance basée sur la corrélation nous permettant d'éliminer les caracteristiques des séries qui ne sont pas cohérentes.

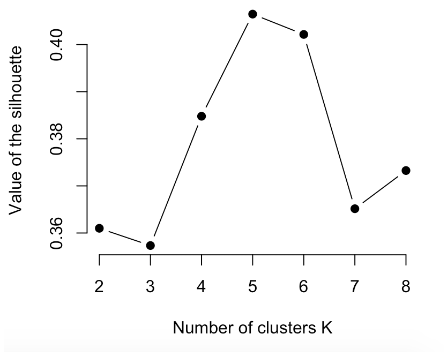

Figure 3: Coefficient Silhouette cor en fonction du nombre de cluster

Le coefficient silhouette est une mesure de similitude d'un objet avec son propre cluster par rapport aux autres clusters. Une valeur élevée indique que l'objet est bien adapté à son propre cluster et mal adapté aux clusters voisins. Si la plupart des objets ont une valeur élevée, alors la configuration du clustering est appropriée.

Dans le cadre de la distance cor, le coefficient silhouette est maximal pour un nombre de cluster k=5 pour lequel nous obtenons les résultats suivants:

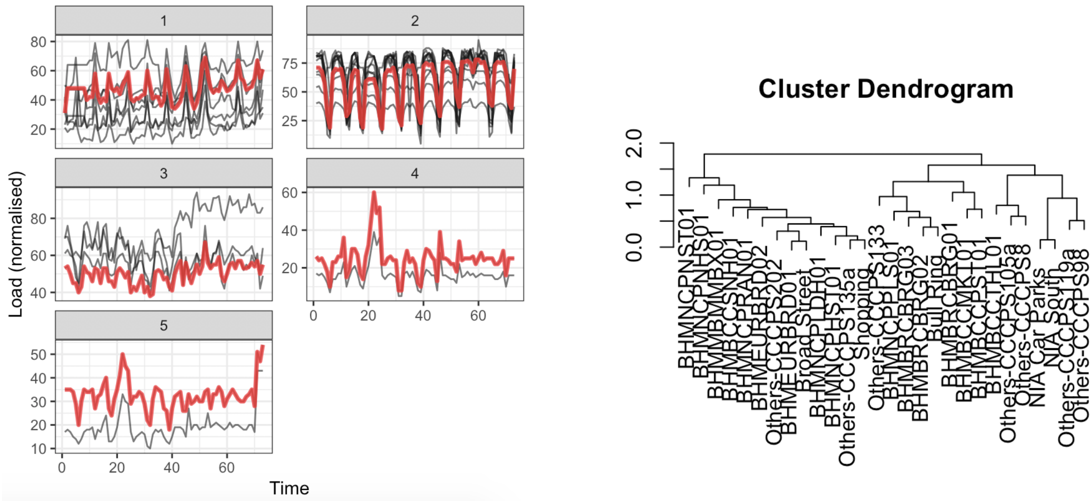

Figure 4: Résultat du clustering Cor: cluster=5

Nous remarquons ici que les deux premiers clusters englobent respectivement 8 et 12 des 28 séries du jeu de données. Les médoides de chaque classe, en rouge nous

permettent d'en apprendre un peu plus sur la structure de chacune des classes: les deux premiers clusters semblent avoir une saisonalité de 7 correspondant aux jours de la semaine tout en étant diamétralement opposés: le cluster 2 semble correspondre aux parkings utilisés le plus souvent au cours de la semaine, probablement pour des utilisateurs se rendant sur leur lieu de travail. Ces parkings sont également très peu fréquentés lors du week-end. À l'inverse, le cluster 1 semble rassembler les parkings dont la fréquentation augmente les week-ends.

Les trois clusters restants peuvent faire figure d'outlier ou de parkings ayant des utilisations spécifiques: par exemple le cluster 4 composé des parkings NIA Car Parks et NIA South sont tout deux des parkings de la salle omnisport de Birmingham, leur forte ressemblance provient donc du fait qu'ils sont situés au même endroit et que leur fréquentation est fortement dépendante des activitées proposées par la salle omnisport: match, concert, athlétisme...

3.1.2 Basé sur la distance DTW

Cependant, la distance basée sur la corrélation n'est pas effective lorsque les séries à classifier sont de tailles différentes, qu'elles ne sont pas synchronisées les unes aux autres en fonction du temps ou bien qu'elles n'ont pas la même amplitude.

La déformation temporelle dynamique est un algorithme permettant de mesurer la similarité entre deux séries qui peuvent varier au cours du temps. Celle-ci mesure l'appariement optimal entre deux séries et tient compte de la forme des séries.

En calculant les distances DTW, les valeurs silhouette nous permettent d'obtenir un clustering en deux groupes:

Les deux clusters sont répartis de manière équitable avec respectivement 12 et 16 séries chacuns. De la même manière que précédement, nos deux clusters semblent se répartir entre les parkings dont la saisonalité est telle que leur occupation est plus fréquente sur une petite perriode comme le week-end pour le premier et une plus grosse comme les jours de la semaine travaillés pour le deuxième.

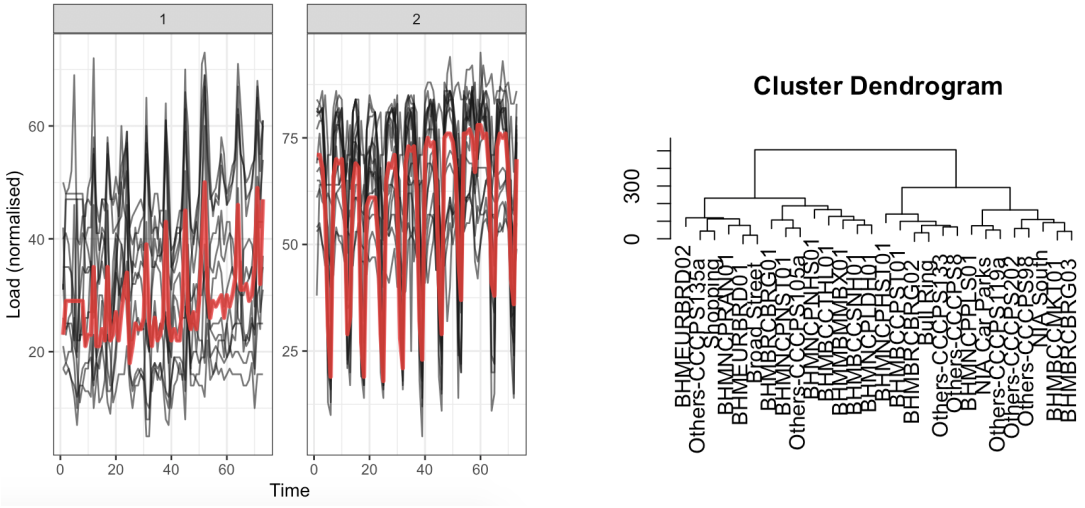

Figure 5: Résultat du clustering DTW: cluster=2

3.1.3 Basé sur les coefficients ARIMA

Nous avons souhaité essayer une dernière méthode de classification de séries temporelles basée sur le modèle ARIMA: La nature intrinsèque d'une série chronologique est généralement telle, que les observations sont dépendantes ou corrélées les unes aux autres. Les processus de moyennes mobiles autorégressives constituent une classe très générale de modèles paramétriques utiles pour décrire ces corrélations. L'idée derrière la classification basée sur le processus ARIMA est de comparer les paramètres du modèle qui correspond le mieux à chaque séries aux autres. Il existe différentes métriques capables de réaliser ce travail, nous avons selectionné la distance de Picolo.

En calculant les distances basées sur les coefficients ARIMA de chaque séries, les valeurs silhouette nous permettent d'obtenir un clustering en quatre groupes:

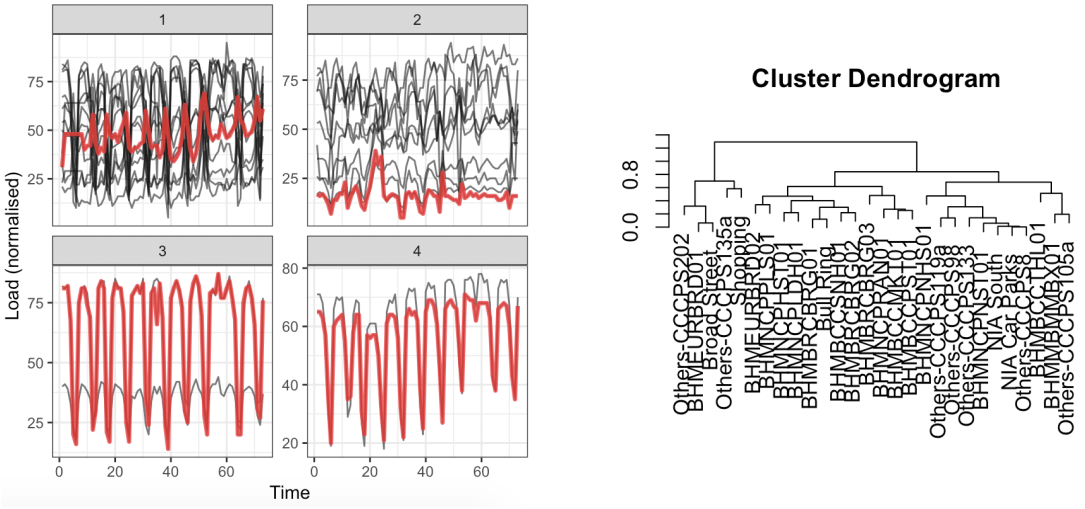

Figure 6: Résultat du clustering Picolo: cluster=2

Les deux derniers clusters peuvent être apparentés à des outliers, ceux-ci étant sous-représentés aux deux premiers cluster mais semblent capturer des séries de saisonalité 7. Le medoide du groupe 2 semble capturer la majorité des parkings dont la variance est assez imprévisible ou du moins, moins que ceux du cluster 1 qui semblent être très réguliers dans leur taux d'occupation en fonction du temps: cela est probablement du à la distance de Picolo, basée sur la distance entre les coefficients ARIMA, ces deux premiers clusters doivent probablement être départagés en fonction de l'apport des AR ou des MA de chaque séries: les séries très "stables" et faciles à modéliser grâce à leur valeur précédentes semblent se regrouper dans le cluster numéro 1.

3.2 Evolution hebdomadaire

Afin de ne plus capturer ces saisonalités hebdomadaires des parkings et de comprendre l'évolution de l'utilisation des parkings dans le temps, nous avons décidé de faire un clustering sur les moyennes hebdomadaires du taux d'occupation de chaque parking: nous permettant alors de plus capturer les tendances de séries en fonction du temps et non leur saisonalité.

3.2.1 Basé sur la corrélation

NIA South, NIA Car Parks, BHMBCCMKT01, BHMNCPNHS01, BHMNCPNST01 semblent être les parkings les plus uniques en terme d'évolution du taux d'occupation

hebdomadaire.

grâce au coefficient silhouette basé sur la corrélation, nous trouvons un nombre de cluster de 2:

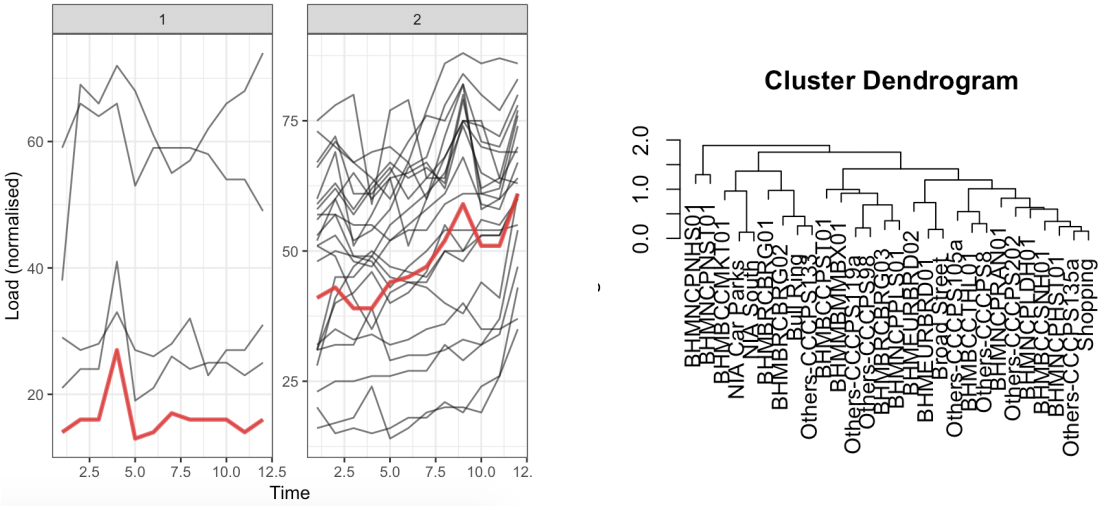

Figure 7: Résultat du clustering cor: cluster=2

Les clusters ici semblent capturer deux types de parkings: le premier ne suivant pas une évolution du taux d'occupation significative mais pour autant subissant certains pics d'activités, liés aux évenements proches de ces parkings: typiquement les parkings proches des centres d'activités comme le complexe sportif. À l'inverse, le cluster 2 semble capturer les parkings dont la tendance est à la hausse et qui nécéssiteraient dans le futur une augmentation de leur places disponibles afin de ne pas être complets: par exemple le parking de Bull Ring, centre commercial de Birmingham dont l'occupation augmente au fil du temps.

3.2.2 Basé sur la distance DTW

De la même manière, notre clustering sur le taux d'occupation hebdomadaire basé sur la distance DTW nous permet d'arriver à la même conclusion que précédement: des parkings dont la tendance est à la hausse et d'autre dont l'utilisation est plus variante.

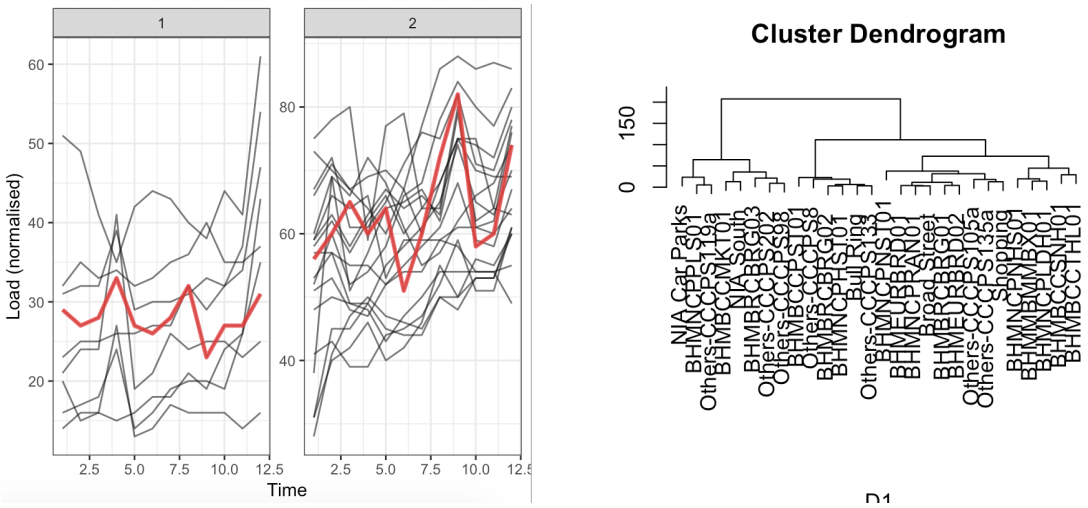

Figure 8: Résultat du clustering DTW: cluster=2

3.2.3 Basé sur les coefficients ARIMA

En calculant les distances basées sur les coefficients ARIMA de chaque séries, les valeurs silhouette nous permettent d'obtenir un clustering en huit groupes:

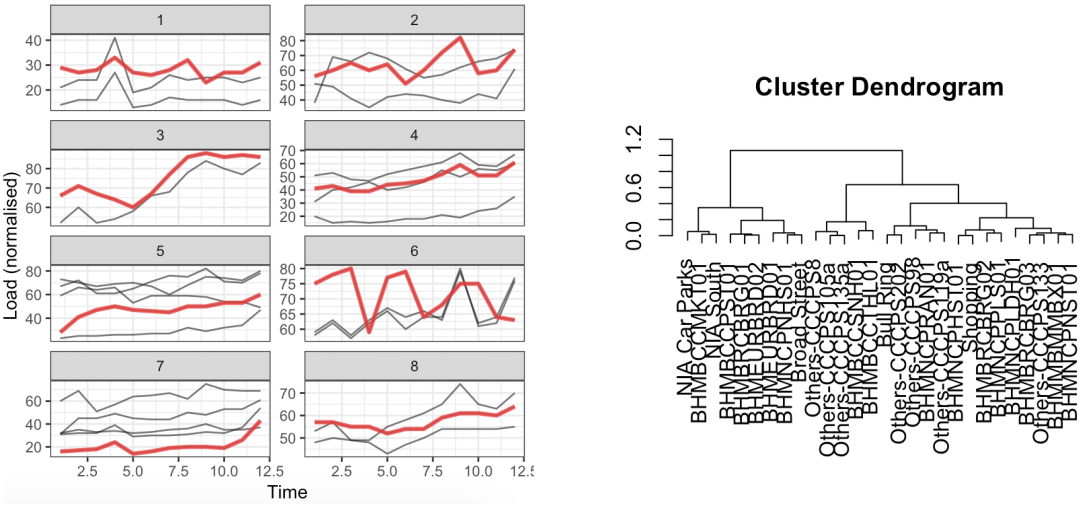

Figure 9: Résultat du clustering Picolo: cluster=8

De ces 8 groupes, certains d'entre eux semblent capturer les parkings dont les taux d'occupation sont assez faibles: par exemple, le groupe 1 est composé des parkings

dont le taux d'occupation n'est jamais superieur à 40% et dont la tendance est stable, ils ne représentent donc pas un risque important à l'avenir. À l'inverse les groupes 3 et 4 semblent regrouper les parkings les plus critiques: forte hausse de la tendance et forte occupation, ceux-ci semblent être en pleine croissance et nécéssiteraient une augmentation du nombre de places disponibles afin de ne pas être complets.

4 Conclusion

Lors de ce projet, nous avons eu l'occasion de manipuler des séries temporelles suivant l'évolution des parkings de Birmingham afin de les regrouper en classes homogènes en fonction de leur taux d'occupation journalier ou hebdomadaire. En fonction du taux d'occupation de ces parkings journaliers, il en ressort une forte distinction entre les parkings occupés majoritairement lors de la semaine et les parkings dont l'évolution augmente très fortement le week-end, cela nous permet donc de situer le taux d'occupation de chacun des parkings en fonction du jour de la semaine. En fonction du taux d'occupation de ces parkings hebdomadaire, il en ressort un groupe de parking dont le taux d'occupation augmente fortement au fil des semaines, et donc qui nécéssiteront à terme d'augmenter leur places disponibles afin de ne pas être complets, et d'autres parkings dont l'utilisation augmente moins au fil des semaines mais dont l'occupation peut parfois subir des "pics" d'occupation lors de jours très précis, typiquement les parkings du stade de Birmingham lors d'un soir de match.

5 Annexe: Modèles de mélanges gaussiens

Afin de parfaire notre analyse, et nos connaissances sur l'apprentissage non supervisé, nous avons souhaité réaliser un clustering de ces mêmes séries via les modèles de mélanges gaussiens: ceux-ci n'étant pas explicitement développés lors de ce cours malgré la présentation de certains cas particuliers comme k-means, les k-médoides ou encore les nuées dynamiques, nous avons jugé préférable de proposer cette analyse en annexe.

Un modèle de mélange gaussien est un modèle statistique exprimé selon une densité mélange. Il sert usuellement à estimer paramétriquement la distribution de variables aléatoires en les modélisant comme une somme de plusieurs gaussiennes. Il s'agit alors de déterminer la variance, la moyenne et l'amplitude de chaque gaussienne. Ces paramètres sont optimisés selon un critère de maximum de vraisemblance pour approcher le plus possible la distribution recherchée.

5.1 Analyse journalière

Nous avons premièrement testé tous les modèles possibles pour toutes les structures de covariance:

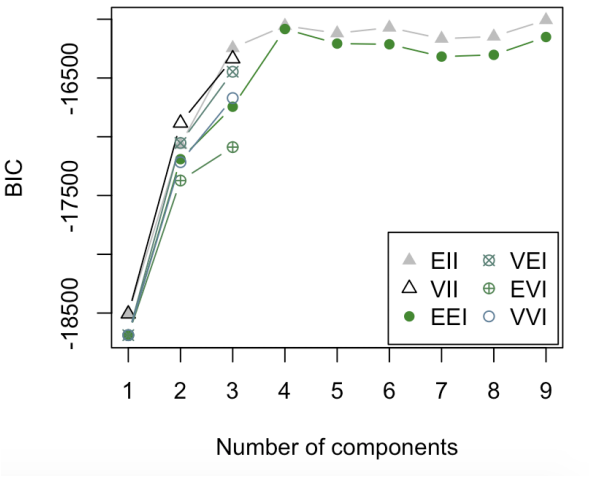

Figure 10: Evolution du critère BIC en fonction du modèle et du nombre de classes

Nous obtenons le clustering suivant en 9 classes:

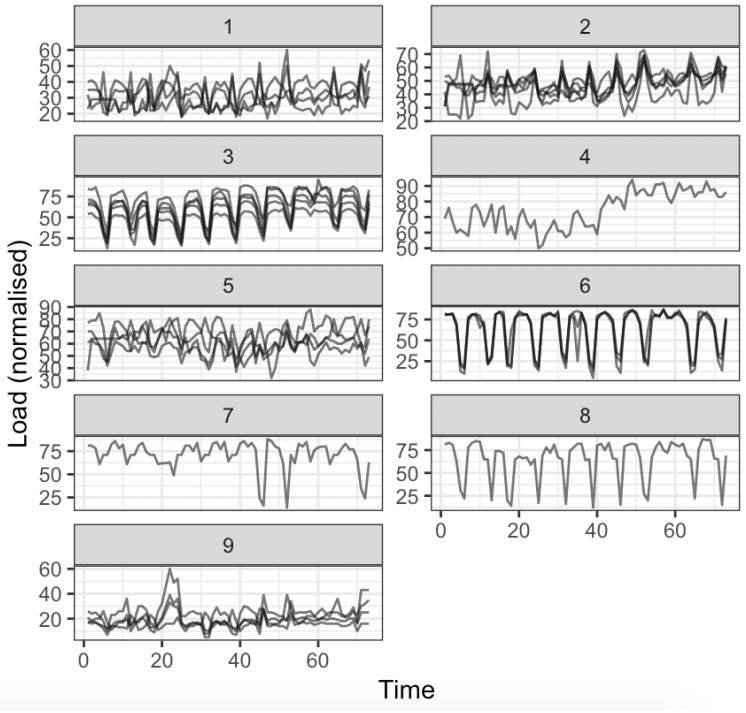

Figure 11: Résultat du clustering GMM

Comme nous l'avions remarqué précédement, certains des parkings sont tellement uniques en terme de taux d'occupation en fonction du temps que le GMM les aura considéré comme des classes à part entière, nous remarquons cependant très bien la présence des groupements dus à la saisonalité en fonction des jours travaillés ou non sur les cluster 1, 3, 6 et 2.

Le cluster 9 semble capturer les parkings subissant des taux d'occupation fortement liés aux évenements sportifs.

5.2 Analyse hebdomadaire

De la même manière, nous avons calculé la valeur du critère BIC en fonction du modèle et du nombre de classe:

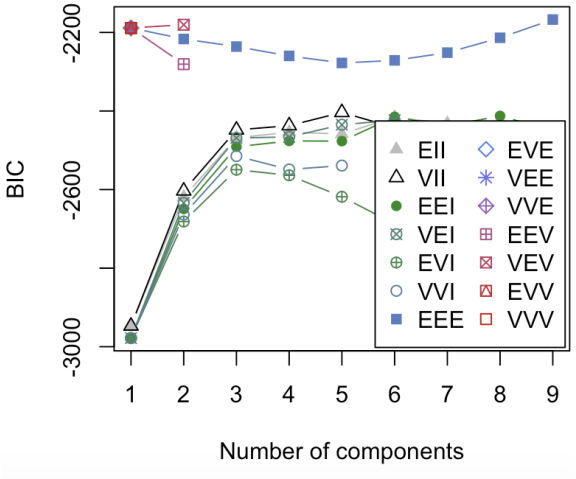

Figure 12: Evolution du critère BIC en fonction du modèle et du nombre de classes

Nous obtenons le clustering suivant en 9 classes:

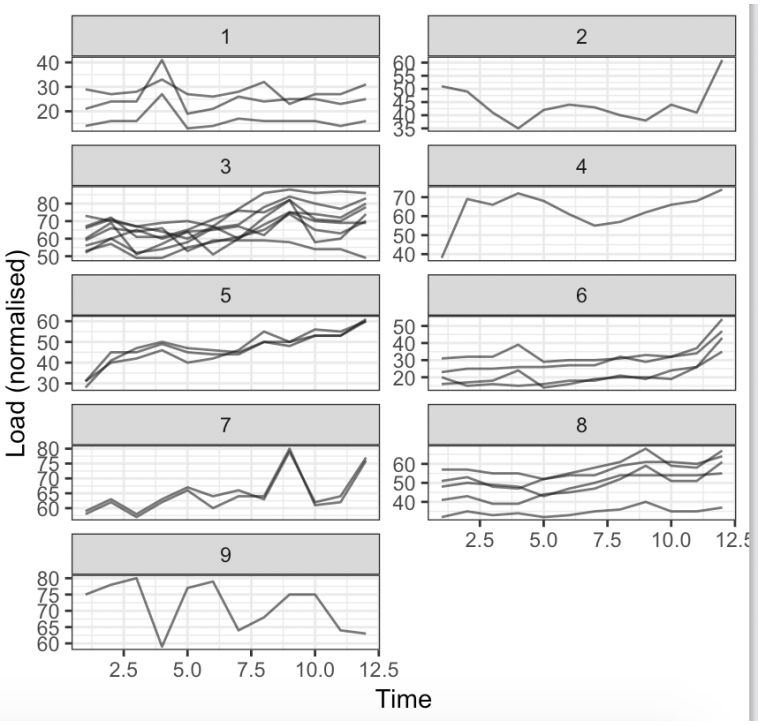

Figure 13: Résultat du clustering GMM

le GMM semble toujours capturer la différence notable entre les parkings dont la tendance est à la hausse des parkings dont l'occupation est variable.

Nous avons souhaité réduire la dimension afin de visualiser les frontières de décisions du modèle, les contours de chacune des classes et leur densité:

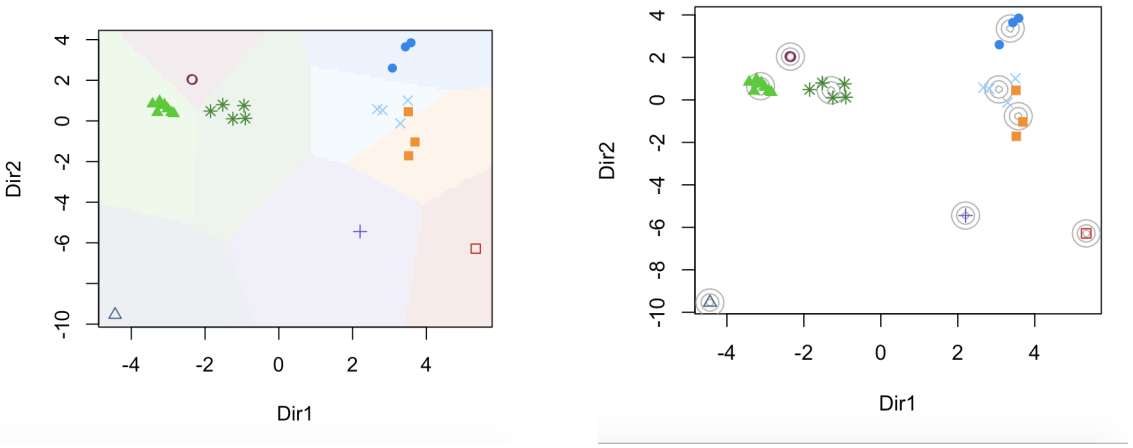

Figure 14: Frontières de décision et contours

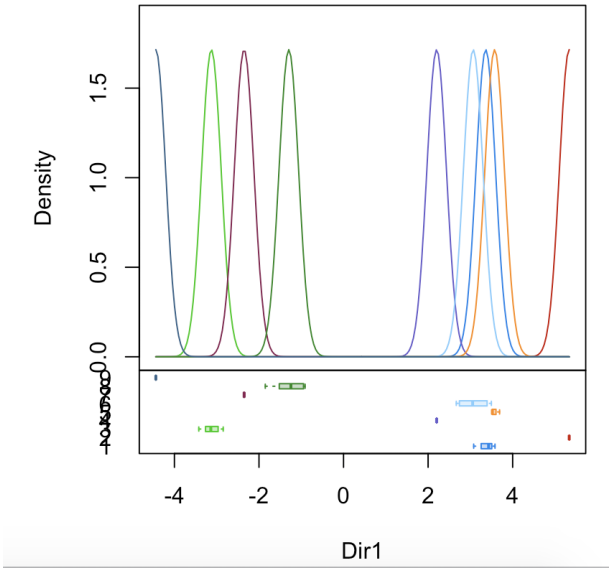

Figure 15: densité des classes du clustering GMM

La densité de chacune des classes nous permet d'avoir une première approche sur les possibles conflits interclasse afin de voir lesquelles sont les plus proches en terme de distribution et donc là ou le modèle pourrait faillir: typiquement ici, les classes sont en règle générale bien séparées, il semble cependant il y avoir un effet de chaine entre deux classes: 1 et 6.