

Projet d'Analyse de données

JIBIDAR Blessing, MOUGNIBAS Théo

Contents

1	Contenu du jeu de données :	1
2	Analyse unidimensionnelle :	2
2.1	Expression des gènes lors des traitements T1,T2 et T3	2
2.2	Expression relative des gènes mesurées à intervalle régulier	2
2.3	Boxplots des différentes variables quantitatives du jeu de données	3
3	Analyse bi-dimensionnelle	3
3.1	Boxplots par paire de variables (qualitative,quantitative)	3
3.2	matrice de covariance des variables quantitatives	5
3.3	Rapport de corrélation Eta2	6
3.4	tables de contingence pour les variables qualitatives, 2 à 2	6
4	Analyse en composantes principales où les Tt sH Rr sont les individus décrits par les gènes.	6
4.1	Interprétation globale du couple de graphes Figure 6	7
4.2	Clustering	8
5	Analyse des gènes	15
5.1	Generation de dataExpMoy	15
5.2	ACP de DataExpMoy	15
5.3	Clustering	17
5.4	Approfondissement de L'ACP avec un clustering de 7 classes	22
5.5	Question BONUS	24

date: "2024-12-04"

1 Contenu du jeu de données :

Le jeu de données contient 3 variables qualitatives nominales représentant l'expression du gène g dont les modalités sont {"sur","sous","non"}. Chaque variable correspond respectivement à la différence d'expression du gène mesurée à la 6ème heure lors du traitement T appartenant à {T1,T2,T3} en moyenne, sur les réplicats {R1,R2}.

$3 \times 6 + 3 \times 6 = 36$ variables quantitatives continues représentant les effets des traitements T1 T2 et T3 sur l'expression des gènes à 1h,2h,3h,4h,5h,6h après l'administration pour les replicats {R1,R2}, par rapport à leur expression à T=0 (sans traitement). Ce jeu de données contient des relevés sur 542 individus, ici des gènes.

2 Analyse unidimensionnelle :

2.1 Expression des gènes lors des traitements T1,T2 et T3

Nous avons réalisé 3 histogrammes pour visualiser les effectifs des gènes en fonction de leur expression relative moyenne à 6h suite aux traitements T1,T2 et T3, cette moyenne est représentée par les variables qualitatives nominales ExpT1,ExpT2 et ExpT3

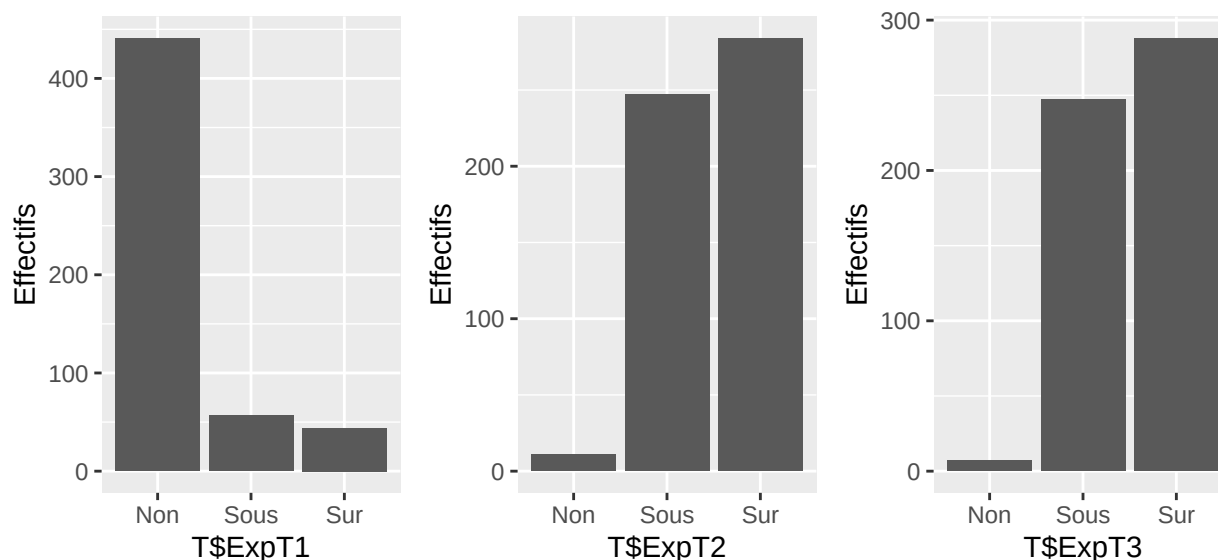


Figure 1: Visualualisation des expressions relative des gènes lors du traitement T1,T2 et T3

2.1.1 Analyse des résultats

Sur la Figure 1 on remarque que les traitements T2 et T3 semblent avoir un effet assez similaire sur l'expression des gènes relevée à la 6ème heure : Une polarisation entre la sous expression et la sur expression qui se partagent presque la totalité des relevés, avec un poids légèrement supérieur à 55% pour la sur-expression au détriment de la sous-expression. Cela a peut être un rapport avec le fait que T3 est une combinaison des traitement T1 et T2.

T1 quant à lui se démarque grandement par une large majorité (Un peu plus de 80%), de gène n'ayant pas changé d'expression après 6h de traitement.

2.2 Expression relative des gènes mesurées à intervalle régulier

Lors d'une analyse supplémentaire des fréquences d'expression des gènes à chaque heure et chaque traitement, nous observons bien une concordance avec l'analyse des expressions des gènes Figure 1. En effet, les histogrammes en rapport avec le traitement 1 sont très nettement regroupés vers 0, soit une expression relative des gènes qui ne change peu. Les histogrammes pour les relevés des variables en lien avec T2 et T3 sont tout aussi similaires aux résultats précédents : La variance de l'expression relative des gènes est plus élevée et on observe bien une polarisation "sous-exprimé-"sur-exprimé" sur les relevés à 6h. Attention, ici on observe aussi que T2 et T3 n'ont pas leur effet caractéristique directement : à 2h, la distribution de

l'expression des gènes semble presque Gaussienne, et à 1h elle ne se distingue pas beaucoup du traitement 1 avec un regroupement sur 0.

c.f le document RMarkdown pour les histogrammes.

2.3 Boxplots des différentes variables quantitatives du jeu de données

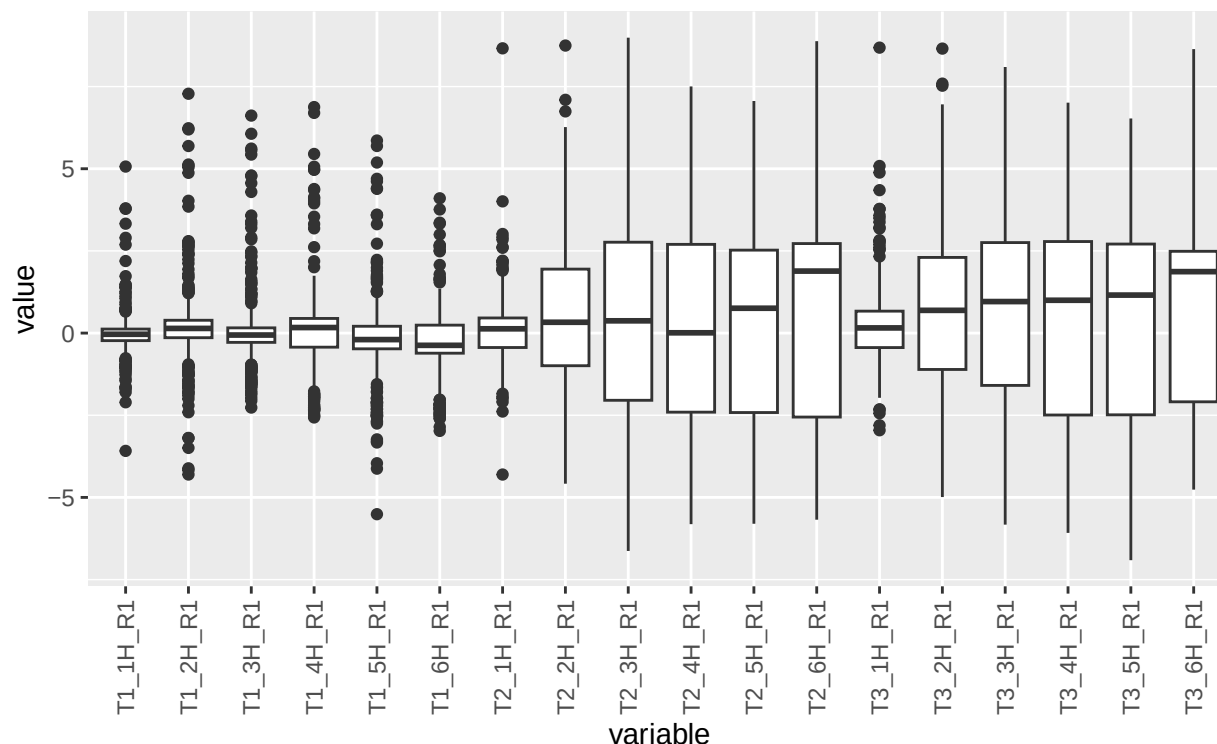


Figure 2: Boxplots des différentes variables du jeu de données lors du réplicat R1

Sur la figure 2 nous avons choisi de n'inclure que le graphe portant sur le réplicat R1 car il est porteur des mêmes informations que celui sur le réplicat R2. L'analyse des boxplots montre que, même sans effectuer de réduction des données, chaque variable présente une dispersion similaire, ce qui suggère qu'il n'est pas nécessaire de procéder à un centrage et à une réduction avant de réaliser l'Analyse en Composantes Principales.

3 Analyse bi-dimensionnelle

3.1 Boxplots par paire de variables (qualitative,quantitative)

Nous avons ensuite générés des graphes contenant des boxplots d'expression des gènes moyennes d'un traitement par rapport à l'expression des gènes à 6h sur un autre traitement. Cela nous a permis d'identifier les changements d'expressions de chaque groupes de gènes sur, sous ou non-exprimés entre un traitement et les deux autres. Les deux réplicats ayant des résultats très similaires, pour ne pas surcharger le rapport nous avons décidé d'afficher seulement le réplicat R1.

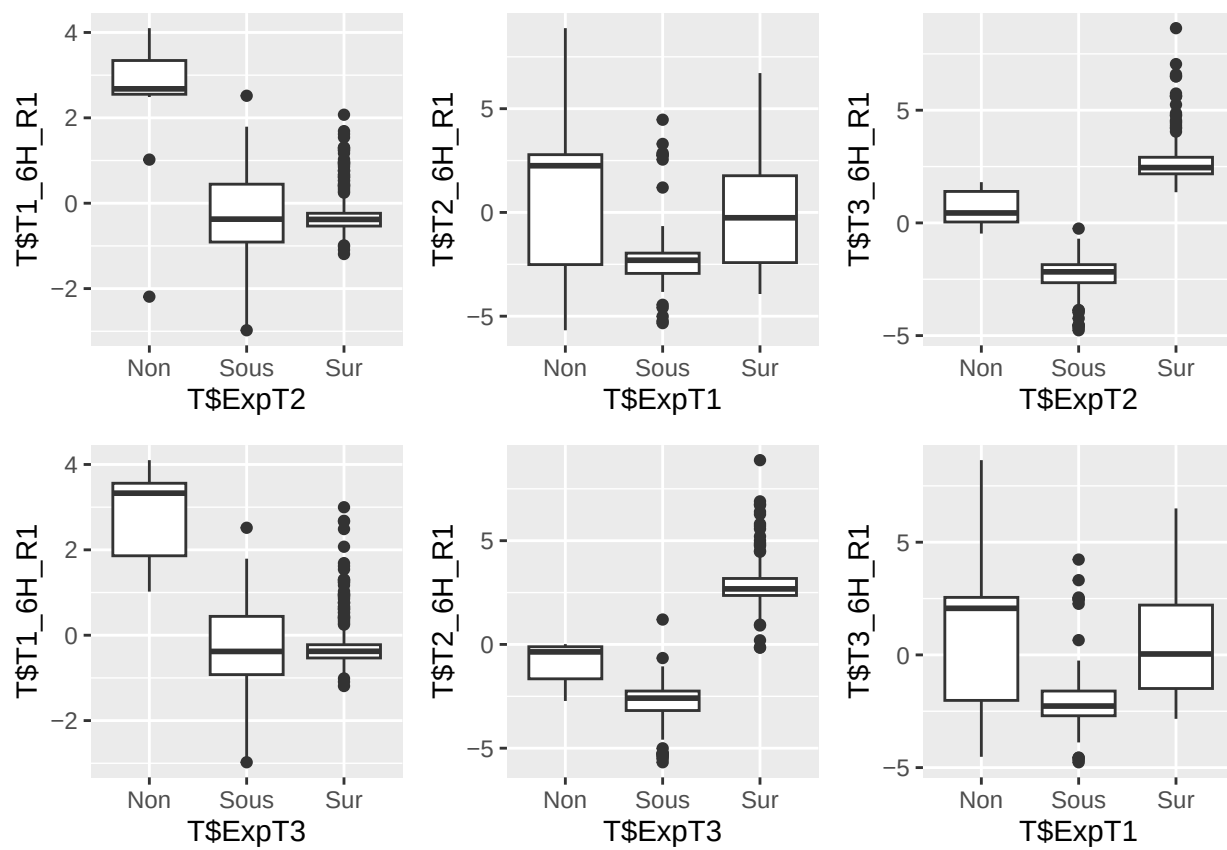


Figure 3: Boxplots par paire de variables (qualitative,quantitative)

3.1.1 Analyse des boxplots :

Nous allons analyser les boxplots obtenus à la Figure 3

3.1.1.1 Traitement 1 Les genes sur-exprimés au T1 sont non-exprimé durant le T2 et T3 (Boxplots T1/ExpT2 et T1/ExpT3) . Il est difficile d'observer une catégorie de genes de T1 qui se soient sous-exprimés ou sur-exprimés dans T2 et T3, ceux qui n'avaient pas changés d'expression relative durant T1, se sont soit sous-exprimé soit sur-exprimé (Boxplots T1/ExpT2 et T1/ExpT3).

3.1.1.2 Traitements 2 et 3 On observe une légère tendance des genes s'étant sous-exprimé avec T1 à se sous-exprimer avec T2 et T3 (Boxplots T2/ExpT1 et T3/expT1). En revanche, il est très clair que T2 et T3 ont les même effets sur les mêmes genes à 6h, toutes les expressions relevées par T2 concordent aux modalités qualitatives moyennes calculées sur T3 (Boxplots T2/ExpT3 et T3/ExpT2).

3.2 matrice de covariance des variables quantitatives

Nous avons généré la matrice de covariance de nos données sans les variables quantitatives. Suite à cela nous avons remarqué qu'il était inutile d'afficher le réplicat R2 dans la matrice et que cela rendait le graphe moins lisible. Nous avons donc décidé de seulement afficher le graphe de la matrice de covariance avec le réplicat R1.

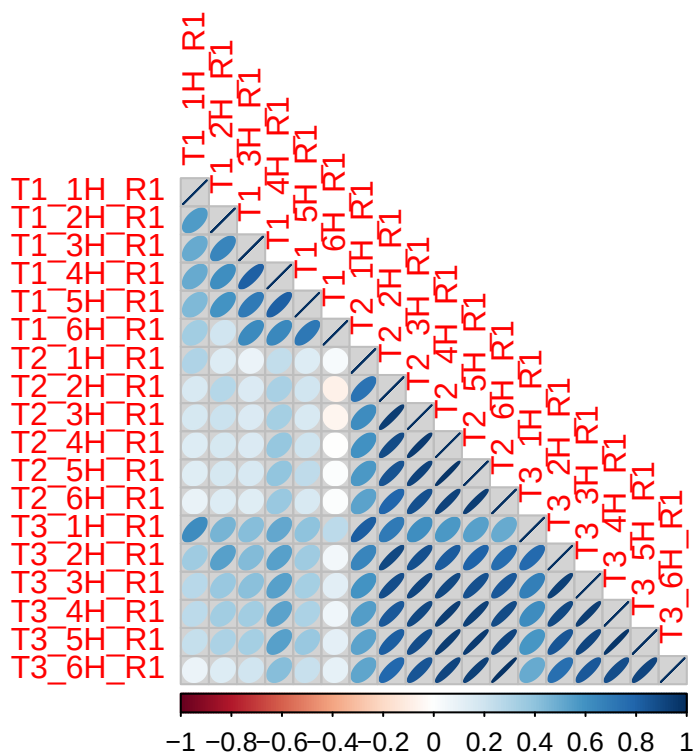


Figure 4: Visualisation de la matrice de covariance des variables quantitatives sur le réplicat R1

3.2.1 analyse de la matrice de covariance

Sur la Figure 4 on observe clairement de grandes zones de de corrélation entre T2 et T3 délimitées par une corrélation plus moyenne à la première heure des deux traitements sûrement dues au fait que T3 est influencé

par T1 aux premières heures. On remarque que T3 est plus fortement corrélé à T1 à 1H que T2 (qui ne l'est que légèrement), semblant indiquer que T1 a un effet avant T2; T3 étant la combinaison des deux traitements.

3.3 Rapport de corrélation Eta2

- Pour la comparaison entre T1 et T2, le rapport de corrélation est de 0,187 pour le réplicat 1 et de 0,199 pour le réplicat 2.
- Pour la comparaison entre T1 et T3, le rapport de corrélation est de 0,152 pour le réplicat 1 et de 0,142 pour le réplicat 2.
- Pour la comparaison entre T2 et T3, le rapport de corrélation est de 0,902 pour le réplicat 1 et de 0,888 pour le réplicat 2. Le calcul du rapport de corrélation η^2 confirme bien notre observation de la grande similarité d'expression des gènes traités avec T2 et T3 et la dissimilarité des expressions des gènes lorsque la plante est traitée avec T1 comparée à T2 et T3, chose normale au vu du peu de gènes affectés par T1.

3.4 tables de contingence pour les variables qualitatives, 2 à 2

```
## [1] "table de contingence entre T1 et T2"
```

```
##
##           Non Sous Sur
##  Non      0  178 263
##  Sous     1   50   6
##  Sur     10   19  15
```

```
## [1] "table de contingence entre T1 et T3"
```

```
##
##           Non Sous Sur
##  Non      0  178 263
##  Sous     0   51   6
##  Sur      7   18  19
```

```
## [1] "table de contingence entre T2 et T3"
```

```
##
##           Non Sous Sur
##  Non      6    1   4
##  Sous     1  246   0
##  Sur      0    0 284
```

Nouvelle confirmation de nos résultats de manière encore plus précise, on observe que T1 ne change pas l'expression de la très grande majorité des gènes. Plus finement, on peut confirmer l'observation faite sur les boxplots tendant à dire que le peu de gènes s'étant sous exprimés avec T1 se sont aussi sous-exprimés avec T2 et T3.

La grande valeur des effectifs partiels sur la diagonale de la table de contingence entre T2 et T3 montre bien la similarité de l'effet de ces deux traitements sur l'expression des gènes.

4 Analyse en composantes principales où les Tt sH Rr sont les individus décrits par les gènes.

L'ACP permet d'obtenir une vision synthétique des données en réduisant leur complexité tout en préservant l'information essentielle. Elle facilite l'identification des tendances et profils d'individus, ainsi que la construction de méta-variables, plus simples à interpréter, tout en évitant les redondances entre les variables. Pour faire cela, nous devons transposer la matrice de données originale qui elle décrivait les gènes (individus) en

fonction des Tt sH Rr. Nous décidons de faire directement une ACP car, comme mentionné plus tôt lors de l'analyse de la Figure 2, nous n'avons pas besoin de centrer et réduire le jeu de données.

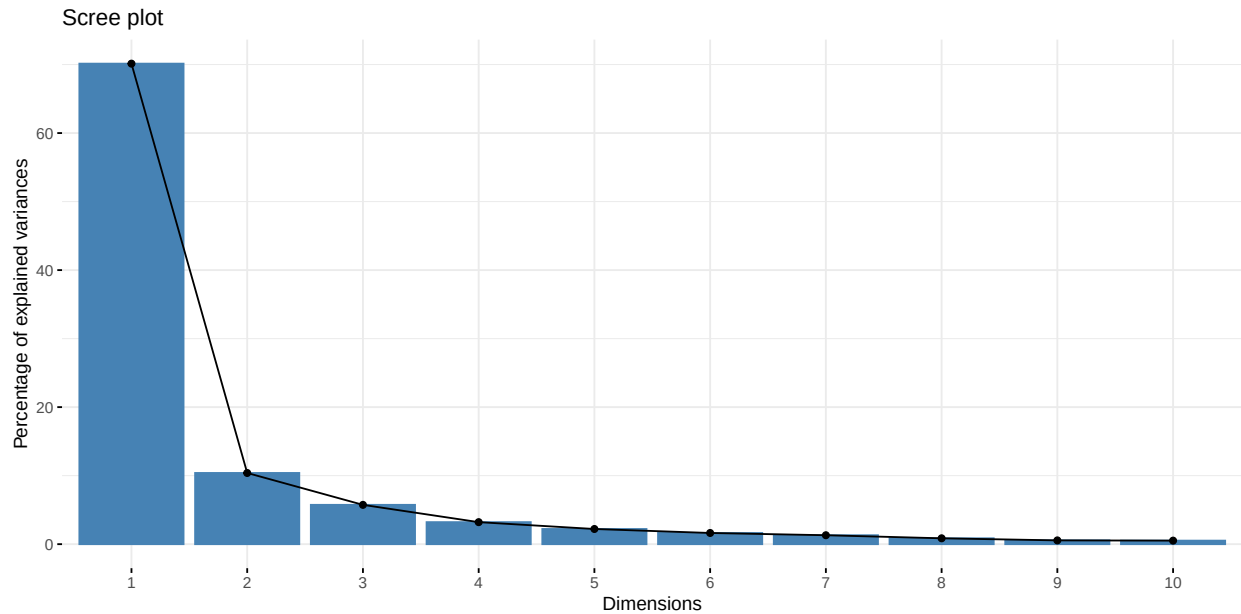


Figure 5: Participation des chaque valeur propre de la matrice de corrélation à l'inertie totale des données

La Figure 5 représente la participation de chaque valeur propre de la matrice de corrélation du jeu de données dans l'inertie totale. L'inertie totale étant la somme des valeurs propres (qui elles sont les inerties axiale associées à l'axe de vecteur directeur le vecteur propre associé), chaque valeur propre est donc une fraction de l'inertie totale. Nous voyons un dépassement des 80% de l'inertie totale rien qu'avec les deux premières valeurs propres, nous en prenons donc les vecteurs propres associés comme axes principaux de l'analyse.

Contexte de la figure 6 : les relevés aux heures sont décrits par les gènes (les gènes sont considérés comme les variables).

- les genes proches d'un axe sont très représentés par celui-ci
- les genes dont l'angle entre eux est petit sont corrélés entre eux
- plus le point représentant un gène est éloigné de l'origine du repère selon un axe, plus ce gène contribue à l'inertie de cet axe

4.1 Interprétation globale du couple de graphes Figure 6

On voit que les genes se polarisent principalement sur l'axe 1 dans un sens ou l'autre. Certaines flèches sont d'une longueur presque du rayon du cercle, indiquant une participation très forte des genes dans la variance expliquée par ces dimensions. Dans chaque "polarité" de fleches selon l'axe 1, il y a des fleches dont la direction est negative d'autres positive selon l'axe 2. Bien que l'on dénote une quantité plus grande de genes corrélés positivement à la dimension 2.

Grâce à l'habillage des graphes des variables (gènes) à l'aide des variables qualitatives, on remarque que les genes se polarisent sur la dimension 1 du côté des valeurs negatives les genes s'étant sous-exprimés à 6h en moyenne sur le traitement T2 et du côté des valeurs positives, ceux qui étaient sur-exprimés. Pour

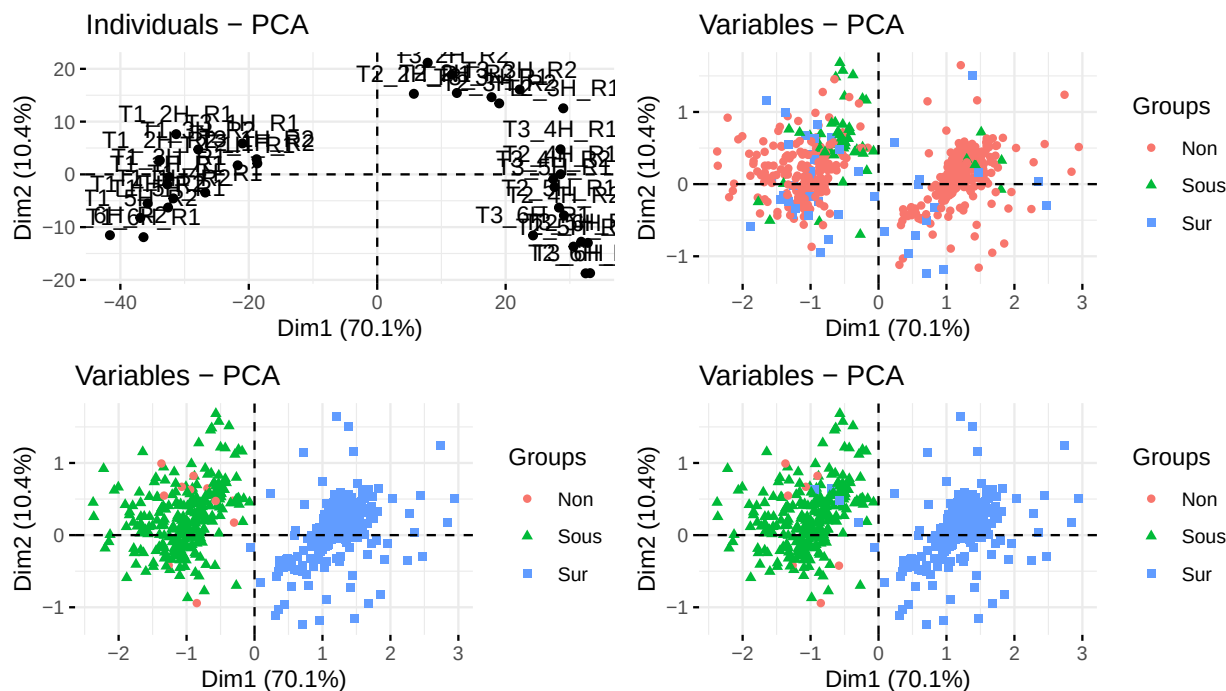


Figure 6: Projection des individus dans le plan factoriel et corrélations des variables avec les composantes principales

la dimension 2, on observe les heures de relevés de traitement s'étaler selon cet axe, on suppose alors que les gènes corrélés à l'axe 2 très positivement s'exprime tôt, et inversement pour ceux qui s'expriment aux environs des 6h.

Le traitement 1 est entièrement groupé sur des valeurs très négatives de l'axe 1. On remarque dans ce groupement la présence des T2 et T3 à la première heure de relevés d'expression des gènes.

Pour le traitement 2 et 3, on les retrouve formant 2 groupements, 1 en haut à droite du graphe contenant les relevés à 2 et 3h puis un groupement s'étalant sur la droite du graphe du centre jusqu'en bas contenant les relevés à partir de 4h.

On en déduit alors que les gènes étant impliqués dans le traitement 1 s'expriment à vitesse moyenne, ni rapidement ni lentement mais qu'ils se sont sous-exprimés dans le traitement T2. De plus, on observe que T2 et T3 visent durant la première heure les gènes qui vont se sous-exprimer et se déplacent petit-à-petit sur le graphe pour toucher les gènes qui vont se sur-exprimer.

4.2 Clustering

L'objectif du clustering est de regrouper des individus (ici, les Tt_sH_Rr) en groupes homogènes selon leurs similarités, sans connaissance préalable des catégories. Cela permet d'identifier des profils ou des comportements communs, facilitant l'interprétation des données et la mise en évidence de structures sous-jacentes.

Nous avons commencé par effectuer un clustering avec la méthode K-means basée sur des estimations de nombre de clusters optimaux avec Silhouette et Calinski-Harabasz ainsi que l'inertie intra-classe, pour finir par effectuer un clustering avec la méthode CAH dont le nombre de classes est déterminé par l'indice de Calinski-Harabasz calculé sur des coupures du dendrogramme.

Pour le dendrogramme, nous avons utilisé la méthode d'agrégation de Ward car il y a trop de points pour utiliser la mesure euclidienne avec efficacité.

4.2.1 Clustering k-means

On commence par afficher l'inertie intra classe en fonction du nombre de classe pour le clustering avec k-means, le but étant de la minimiser tout en gardant un nombre de classes raisonnable.

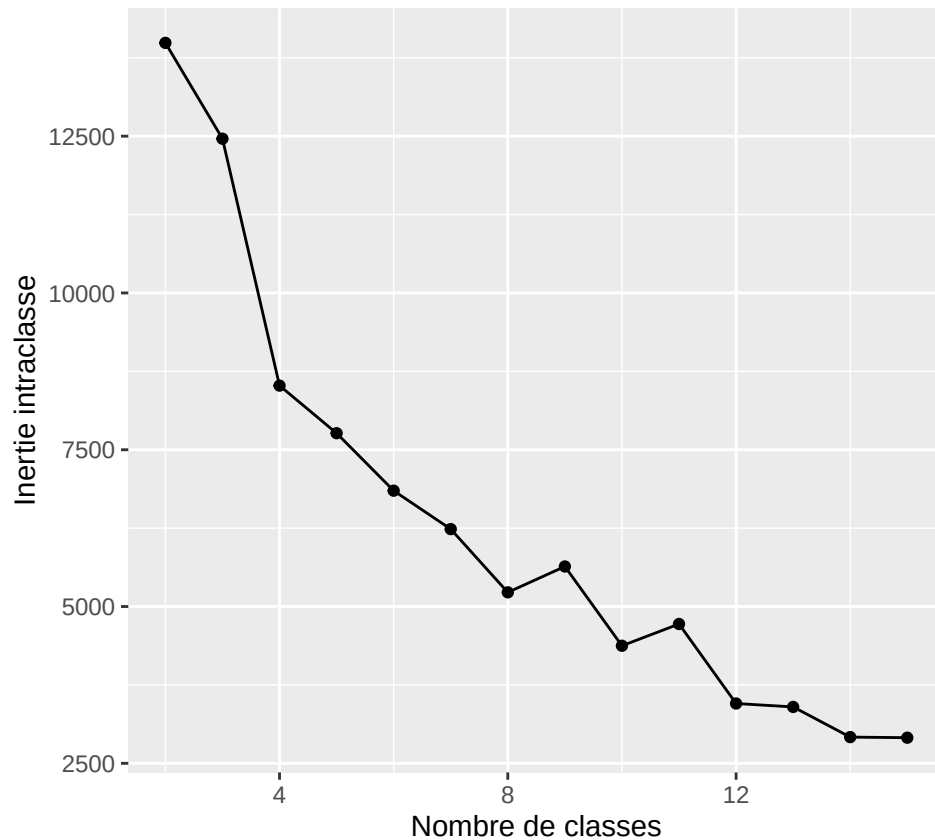


Figure 7: Visualisation de l'inertie intra-classe en fonction du nombre de classes demandées pour le clustering

Sur la Figure 7 on dénote un coude dans la courbe d'inertie intraclasse aux alentours de 4 clusters.

À présent on va afficher sur la Figure 8 l'évolution de l'indicateur Silhouette en fonction du nombre de classe ainsi que l'évolution de l'indice de Calinski-Harabasz en fonction du nombre de classe.

Sur la Figure 8 Silhouette et l'indice de Calinski-Harabasz ont un pic à 2, mais cela est normal sachant que Silhouette et l'indice de Calinski-Harabasz ont tendance à sous-estimer le nombre de clusters.

4.2.2 visualisation du clustering

Après comparaison sur la Figure 9, on choisit 2 classes car cela nous semble plus optimal que 4.

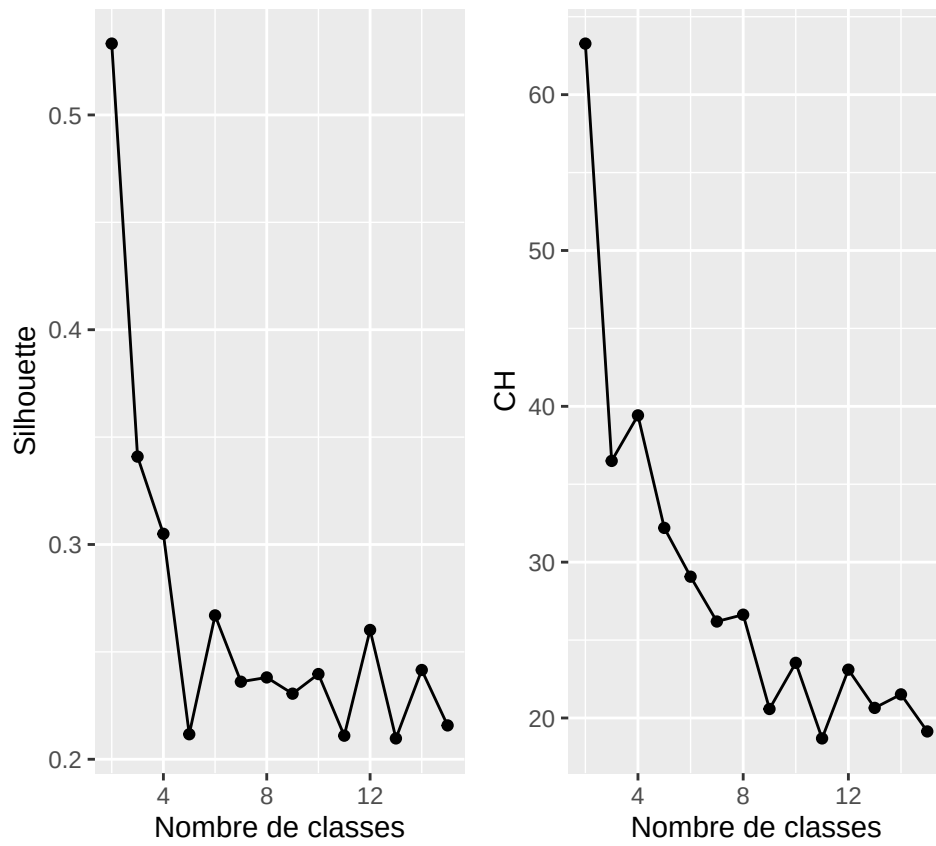


Figure 8: Visualisation du critère de Silhouette et de Calinski-Harabasz en fonction du nombre de classes demandées pour le clustering



Figure 9: Visualisation des clusters générés par la méthode kmeans dans le plan factoriel 1,2

4.2.3 clustering CAH

À présent on va afficher l'évolution de l'indice de Calinski-Harabasz en fonction du nombre de classes utilisées pour découper le dendrogramme.

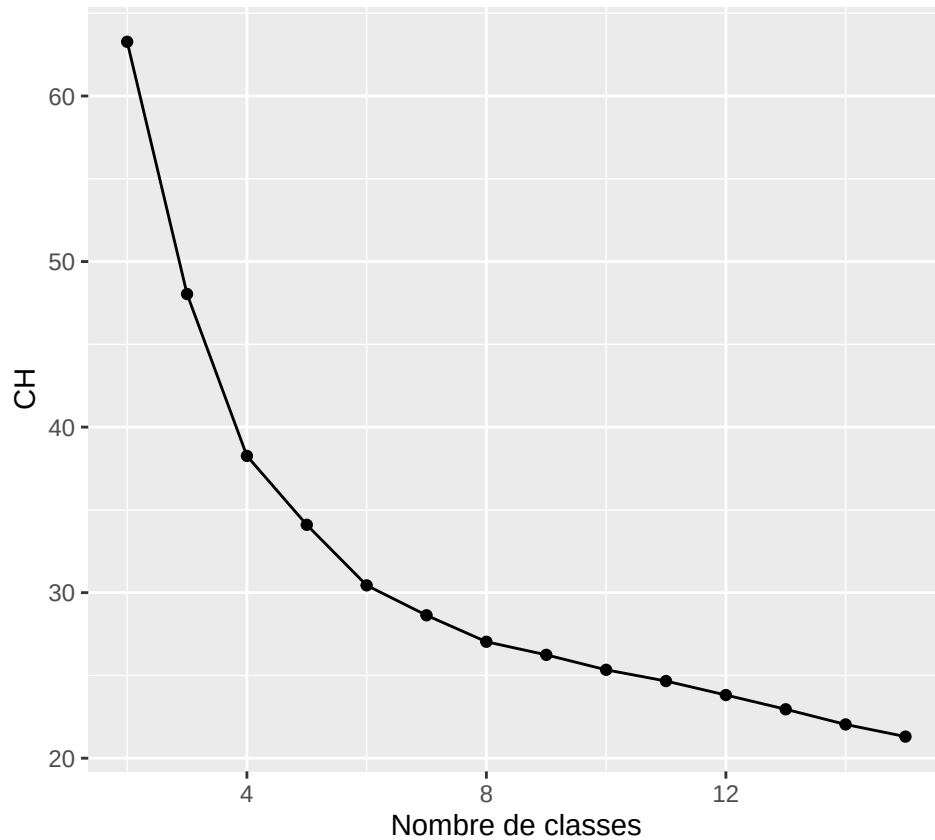


Figure 10: Visualisation du critère de Calinski-Harabasz en fonction du nombre de classes demandées pour le clustering

En regardant la Figure 10, on choisit alors de prendre 2 classes par observation du maximum du graphe.

4.2.4 Comparaison des clusterings

A 2 classes nous obtenons une classification qui ne change pas entre chaque méthode. Nous décidons donc qu'il s'agit donc d'un bon choix de nombre de classes.

La classification obtenue est en accord avec les observations faites lors de l'ACP, on y retrouve plus ou moins les mêmes groupements : celui majoritairement composé des relevés de T1 avec une majorité de gènes sans changement d'expression relative et celui composé des relevés de T2 et T3.

Sur la Figure 12 nous observons que les 2 clustering nous donnent les mêmes clusters peuplés des mêmes gènes.

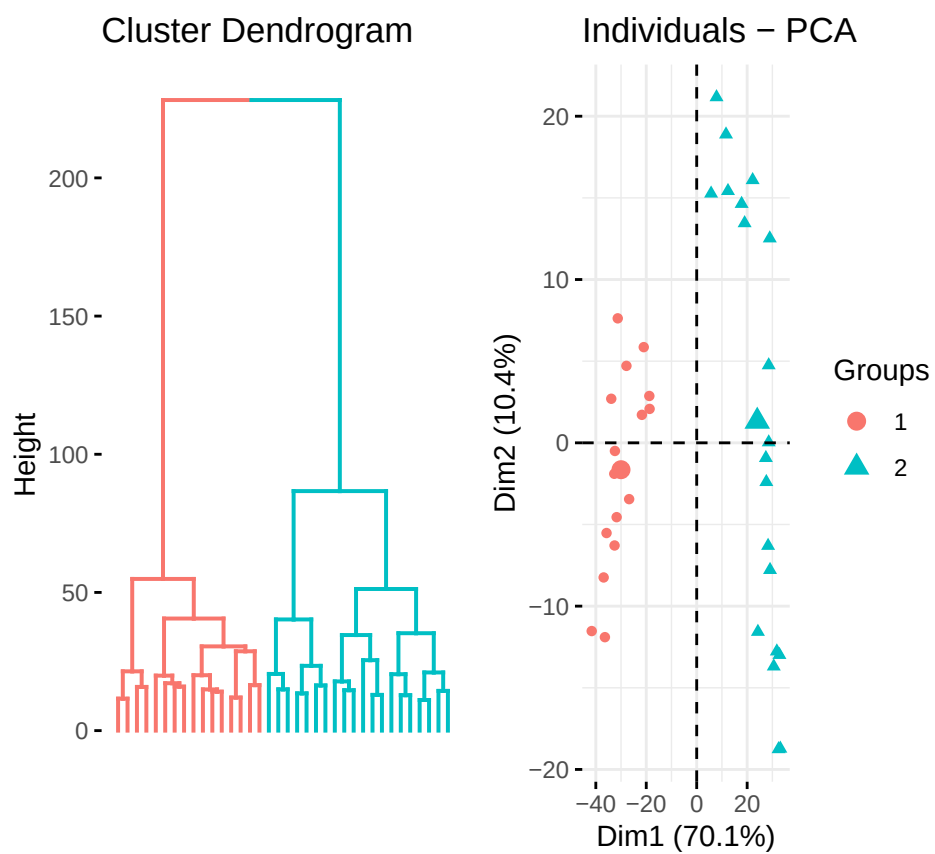


Figure 11: Dendrogramme du clustering de l'ACP des variables Tt en tant qu'individus, obtenu par méthode CAH

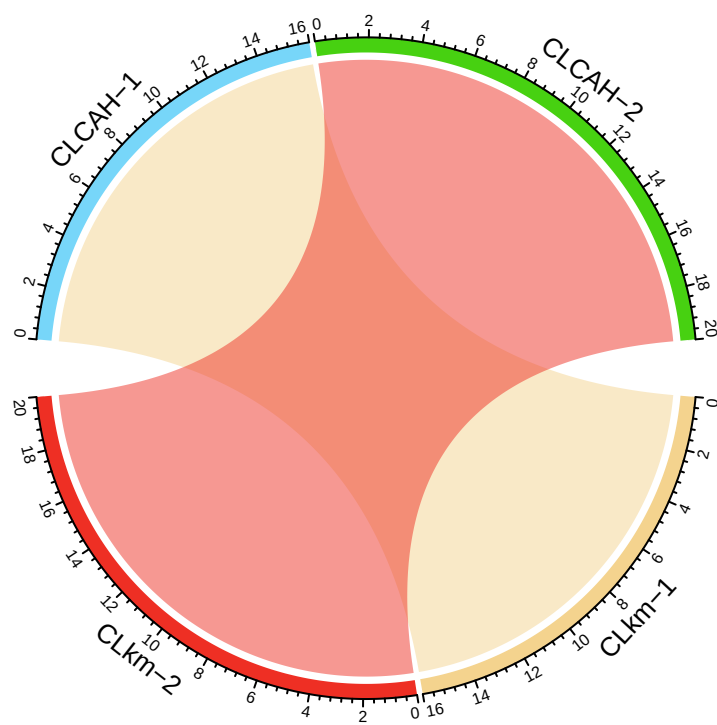


Figure 12: Diagramme mettant en valeur l'exacte correspondance des classes entre les deux clusterings

5 Analyse des gènes

5.1 Generation de dataExpMoy

Nous construisons le jeu de données DataExpMoy contenant la moyenne des expressions sur les réplicats de chaque gène, pour chaque traitement et chaque heure. DataExpMoy est donc une matrice de taille 542×18 .

5.2 ACP de DataExpMoy

Pour l'ACP de DataExpMoy, nous avons centré et réduit les données car cela donnait des résultats sensiblement plus exploitables au niveau des indicateurs pour le clustering.

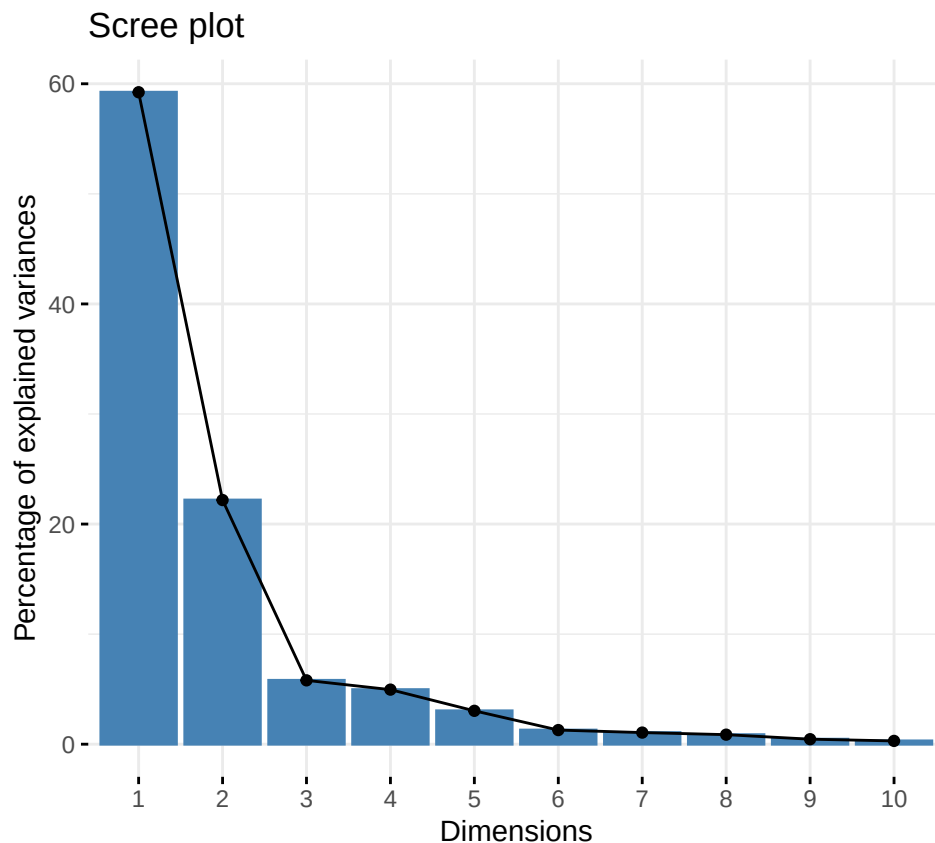


Figure 13: Participation de chaque valeur propre à l'inertie totale des données

Sur la Figure 13 on voit qu'on dépasse 80% de l'inertie totale avec les deux premières valeurs propres, on prend donc les vecteurs propres associés à ces deux valeurs propres comme composantes principales de notre ACP.

En analysant la Figure 14, il semble y apparaître deux groupes principaux de gènes (individus), ce qui pourrait refléter une différenciation claire entre les niveaux d'expression sous différents traitements. Les points plus

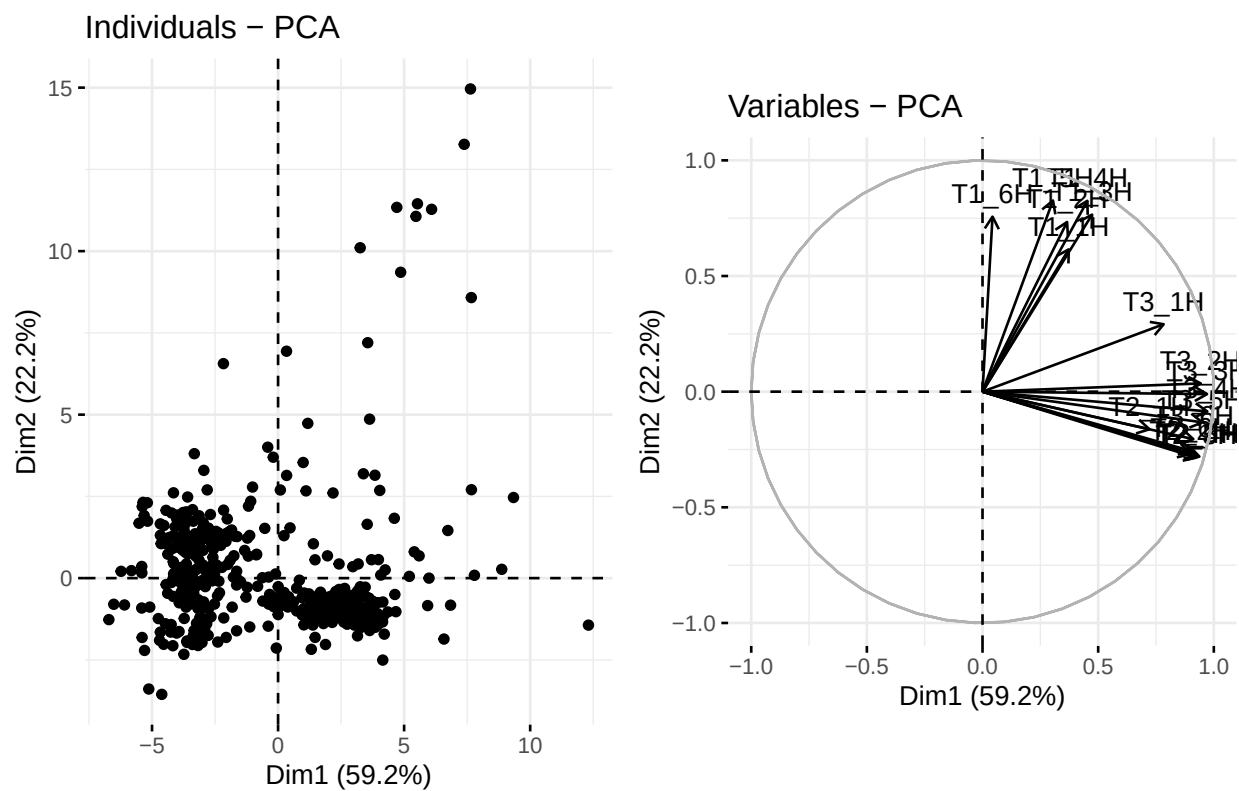


Figure 14: Couple de graphes représentant la Projection des individus sur un plan factoriel (droite) et la Corrélations des variables avec les composantes principales (gauche)

éloignés du centre vers le haut représentent des gènes ayant des comportements plus spécifiques ou atypiques.

Le premier axe (DIM 1) est fortement corrélé avec les traitements T2, T3 à des moments précis, en particulier pour les temps plus avancés (par exemple, T3 4H, T3 5H, T3 6H). Cela suggère que Dim1 reflète les différences d'expression liées à l'effet des traitements T2 au fil du temps, qui a la particularité de faire exprimer ses gènes sur les temps plus longs. Cela explique aussi la corrélation des variables de relevés de T3 à cette dimension, car T3 est un mélange de T1 et T2. Le deuxième axe (DIM2) semble corrélé avec des effets à court terme, typiques de T1. Cela pourrait indiquer des gènes qui réagissent rapidement mais dont l'effet s'atténue à long terme.

En regardant la corrélation des variables qui représentent les relevés sur T1, comme elles sont très positivement corrélées avec la dimension 2, on en déduit que les gènes vers les valeurs élevées de la dimension 2 sont ceux ciblés par le traitement 1, et 3.

Comme T3 est une combinaison des traitements T1 et T2, il semble structurer fortement l'axe principal (Dim1) pour des temps intermédiaires et avancés, mais est notablement corrélé positivement à la dimension 2 pour les premières heures de traitement, correspondant bien avec le fait que T1 commence à faire exprimer les gènes qu'il vise de manière plus rapide que les autres traitements. On retrouve bien le fait que T3 est la somme des deux autres traitements.

5.2.1 Ajout des variables qualitatives

Nous allons maintenant afficher les points dans les dimensions de l'ACP, colorés en fonction des modalités des 3 variables qualitatives ExpT1, ExpT2 et ExpT3. Cela peut nous permettre d'affiner notre interprétation des résultats de l'ACP. Cela nous a donné les graphes Figure 15

5.2.2 analyse

Sur la Figure 15, on observe donc 3 clusters de gènes : - Les gènes étant poussés à la sur-expression par T1, affichés comme non-exprimés durant T2 et partiellement sur-exprimés avec T3 - Les gènes étant sous-exprimés (verts) durant T2 et T3, non-exprimés durant T1. - Les gènes étant sur-exprimés (bleus) durant T2 et T3, non-exprimés durant T1.

On confirme donc bien notre analyse descriptive préliminaire, et nos suppositions de sens prêtés aux dimensions 1 et 2 de l'ACP.

5.3 Clustering

L'objectif de ce clustering est de regrouper des individus (ici, des gènes) en groupes homogènes selon leurs similarités.

En lisant le graphe de l'inertie intra-classe (Figure 16), on observe le coude aux alentours de 3 classes. En observant l'indice de Silhouette, on remarque le pic recherché à 3 classes également. Les indices concordants, on en déduit que le choix optimal du nombre de classes pour le clustering est 3.

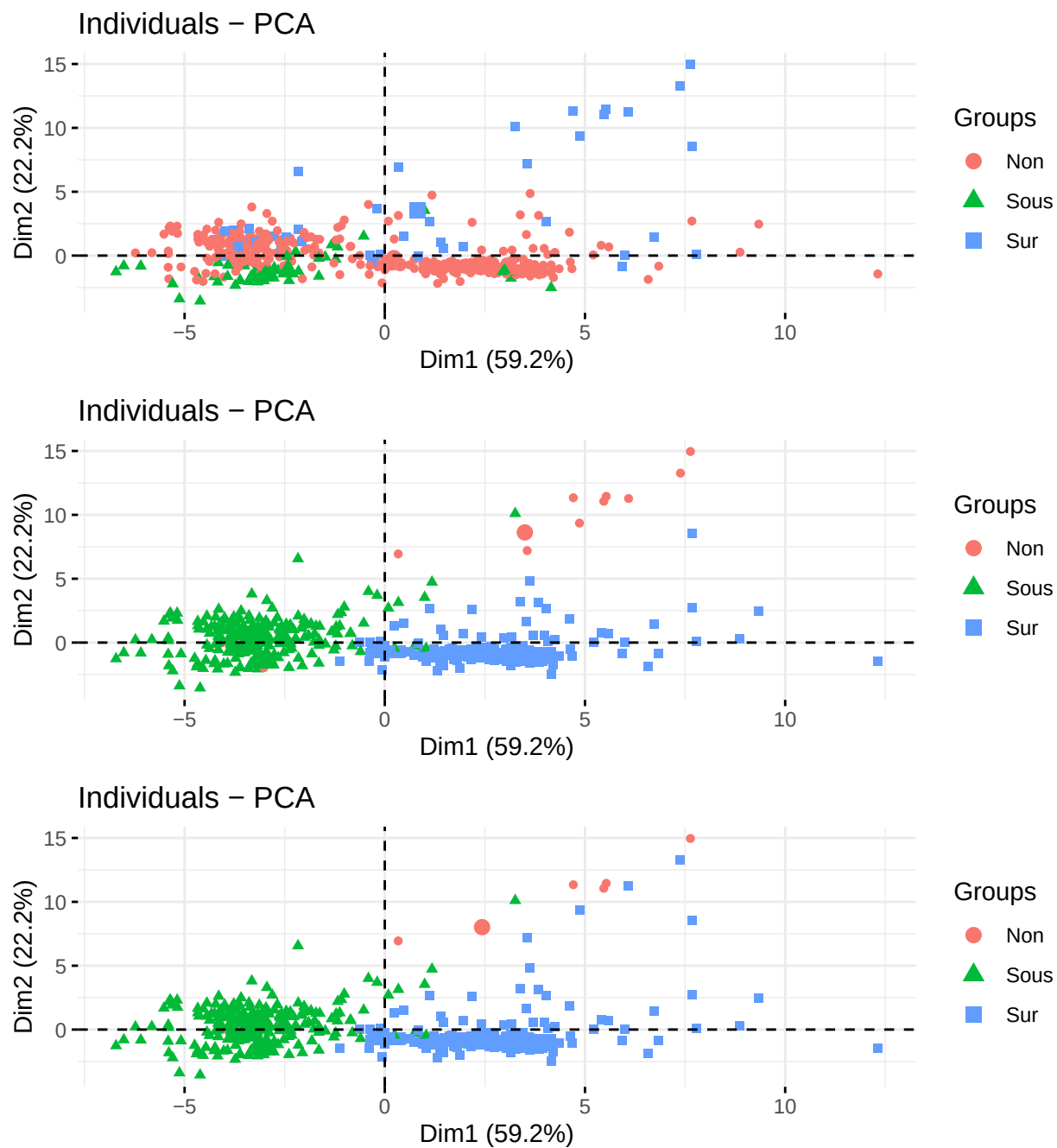


Figure 15: Données sous forme de points dans les dimensions de l'ACP, colorés en fonction des modalités des 3 variables qualitatives ExpT1, ExpT2 et ExpT3.

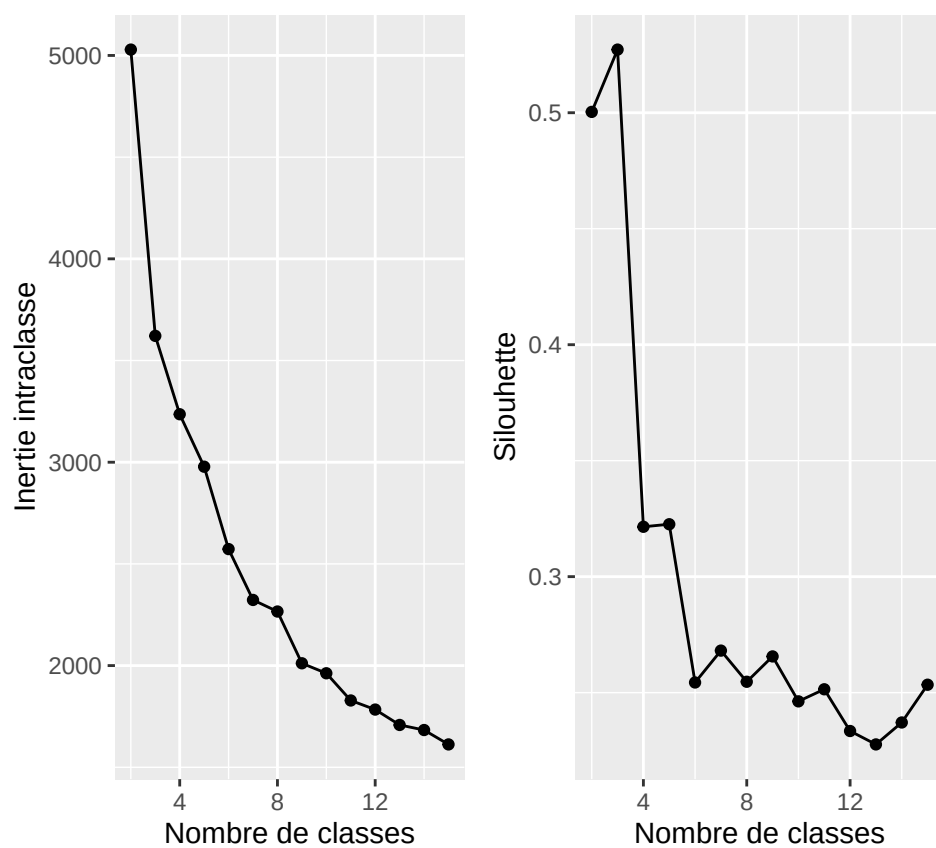


Figure 16: Visualisation de l'inertie intra-classe et Silouhette en fonction du nombre de classes demandées pour le clustering

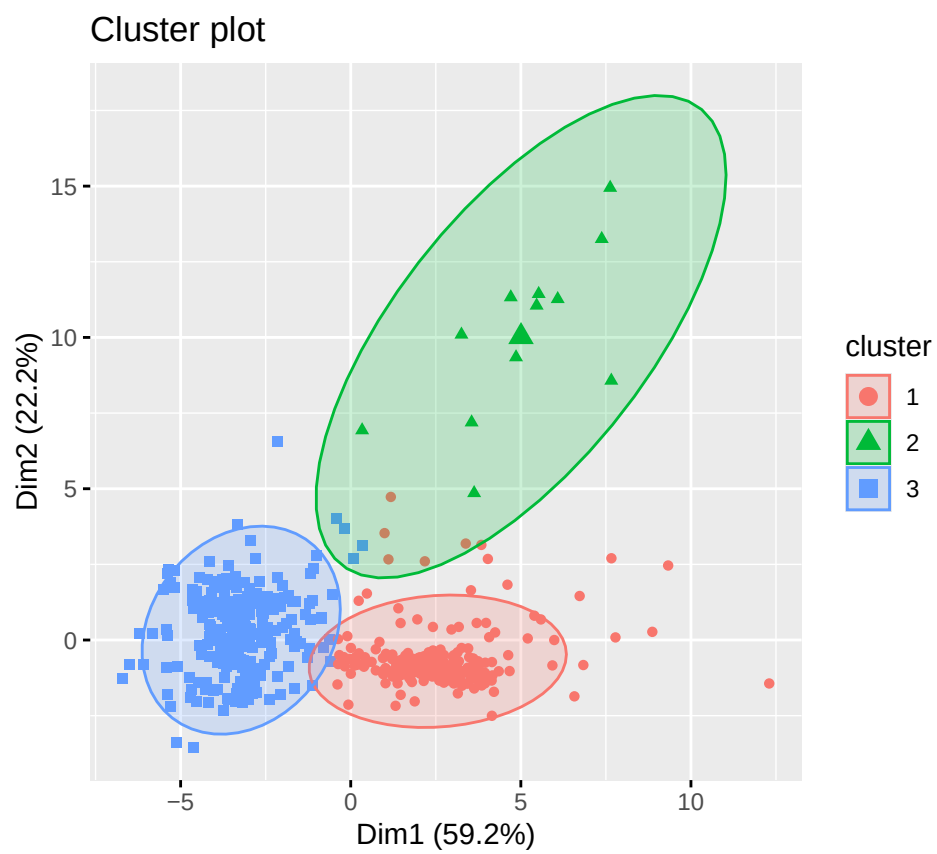


Figure 17: Visualisation des clusters générés par la méthode kmeans dans le plan factoriel 1,2

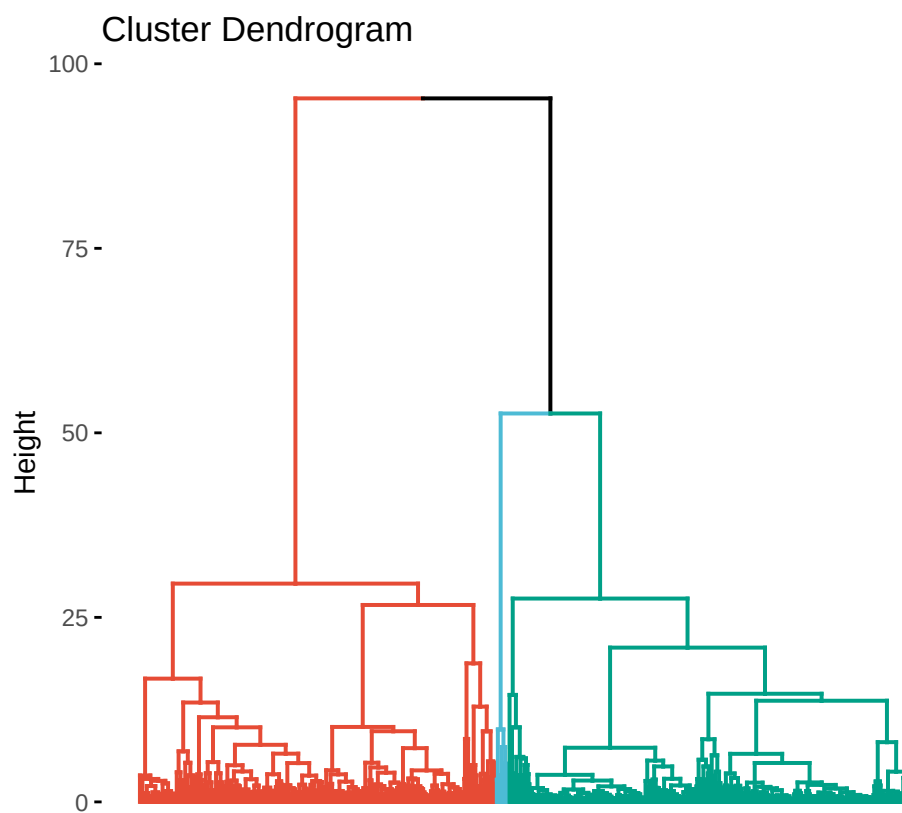


Figure 18: Dendrogramme du clustering obtenu par méthode CAH

5.3.1 analyse des clusterings

On remarque que le clustering de la figure 17 correspond aux groupes que nous avons discernés sur la Figure 15. On en déduit donc la corrélation des gènes de ces classes avec des expressions moyennes similaires dans chaque classe et distinctes entre les différentes classes.

5.3.2 Evolution de l'expression des genes en fonction de leur traitement et cluster

Nous avons généré 3 graphiques (Figure 19) , représentant l'évolution de l'expression moyenne des gènes par cluster et par traitement. Cela pourra nous permettre d'affiner nos observations sur les traitements.

Using GeneID, Cluster as id variables

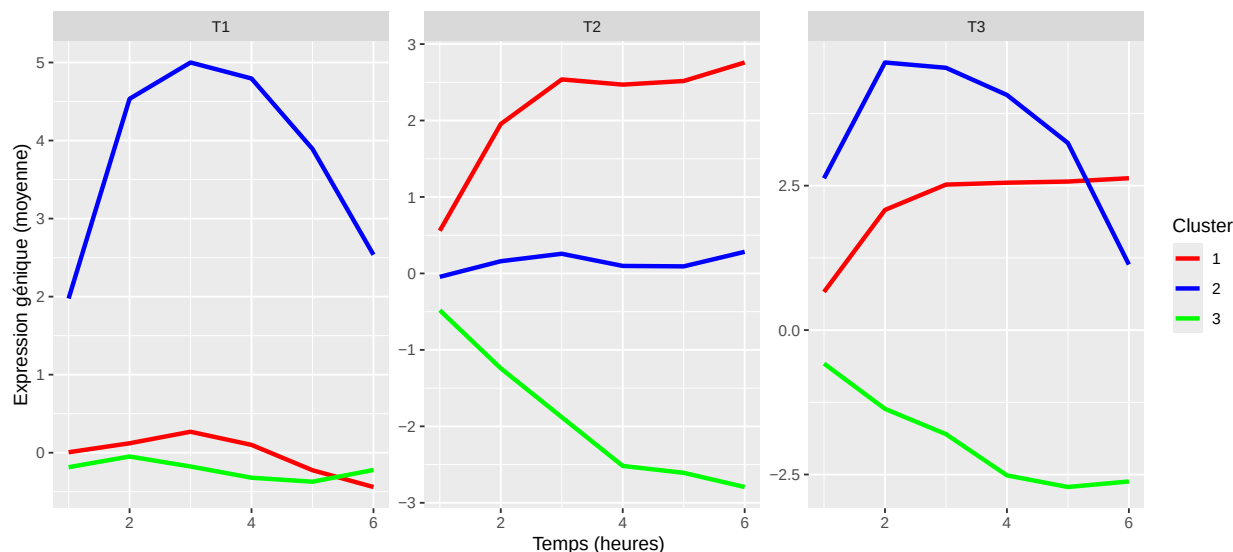


Figure 19: Graphiques représentant l'évolution de l'expression moyenne des gènes par cluster et par traitement

On lis sur les graphiques de la Figure 19 la liaison de toutes les observations que nous avons pu faire :

On remarque sur T1 le fait que le cluster 2 correspond au peu de gènes qu'il influence et qui vont se sur-exprimer plus tôt dans le traitement, pour qu'ensuite l'expression relative re-diminue. Lors du traitement 2, les gènes de ce cluster ne vont pas/peu s'exprimer et lors du traitement 3, ils vont s'exprimer de la même façon que lors du traitement 1, ce qui est logique sachant que T3 est le mélange de T2 et T1.

Comme lors de l'analyse descriptive, on observe les deux autres clusters se sous-exprimer légèrement lors de T1 et se sur et sous exprimer pour T2 et T3, tout en conservant leur expression dans le temps. Le cluster 1 correspond aux gènes sur-exprimés et le cluster 3 aux gènes sous-exprimés pour T2 et T3.

5.4 Approfondissement de L'ACP avec un clustering de 7 classes

Sur la figure 16, sur le graphe de l'inertie intraclasse on peut aussi remarquer un coude aux alentours de 7 classes, nous avons donc décidé de compléter notre analyse par un clustering à 7 classes dont on peut voir la projection dans le plan factoriel sur la Figure 20 et Figure 21 l'évolution de l'expression moyenne des gènes par cluster, en fonction du temps.

Using GeneID, Cluster as id variables

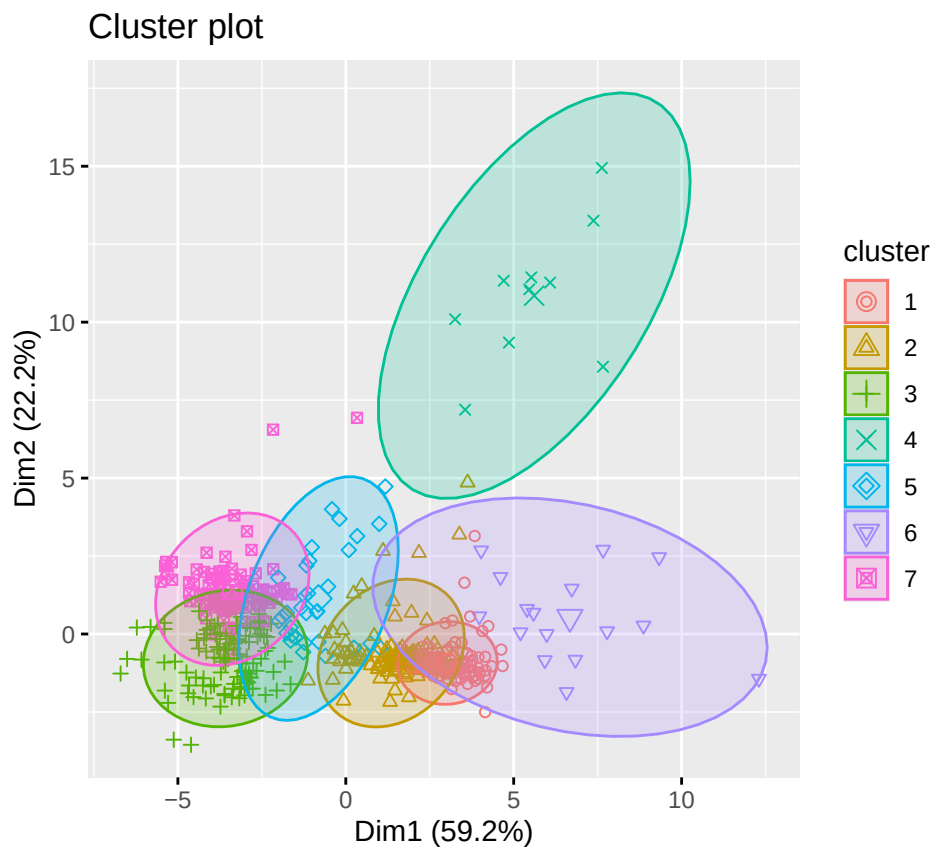


Figure 20: Visualisation des clusters générés par la méthode kmeans dans le plan factoriel 1,2

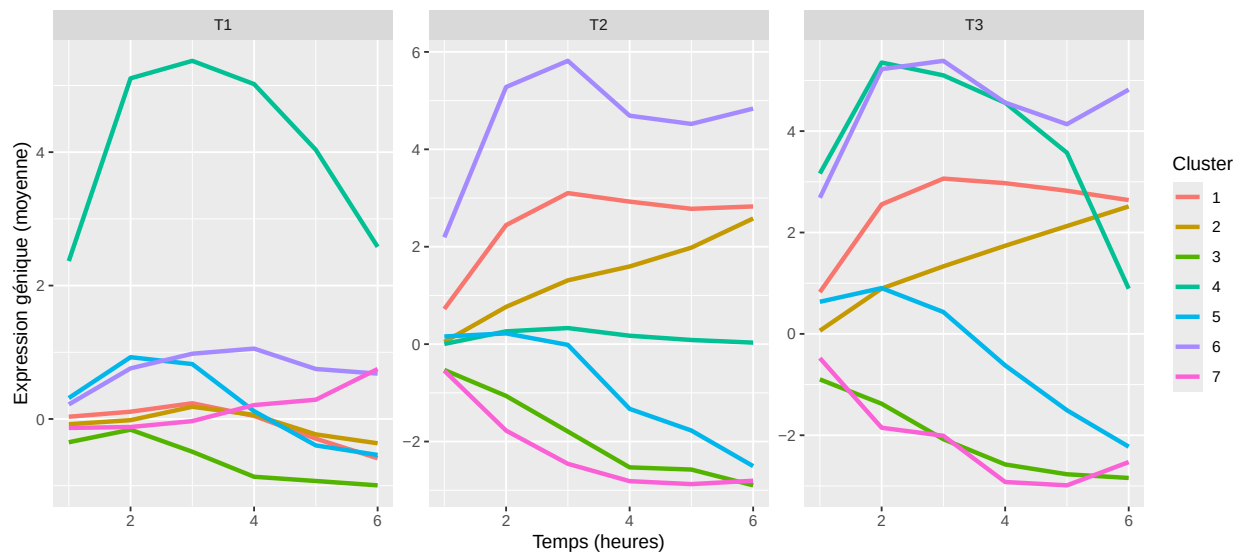


Figure 21: Graphiques représentant l'évolution de l'expression moyenne des gènes par cluster et par traitement

5.4.1 Analyse

Avec plus de clusters, on retrouve plus de précisions dans les comportements différents, avec différentes trajectoire d'expression (et donc comportement entre le premier et dernier relevé) pour les gènes qui finissent sous-exprimés et ceux qui finissent sur exprimés dans T2 et T3 bien que le dernier relevé donne toujours les mêmes tendances globales. On observe par exemple, une classe de gènes qui se sur-exprime, se calme, puis se re-sur exprime. Chose que l'on ne distinguait pas en effectuant les moyennes des comportements sur uniquement 3 clusters : nous avons juste une courbe tendant vers la sur-expression à 6h. Il est tout de même intéressant de noter que la courbe d'expression moyenne pour les gènes sur-exprimés par T1 ne bouge pas par rapport à l'analyse sur 3 clusters, il semblerait donc que ce cluster n'ait sensiblement pas bougé. On peut aussi confirmer les tendances vues dans la première ACP, comme le fait que T1 vise des gènes qui s'expriment à un horizon temporel moyen.

5.5 Question BONUS

Nous n'avons pas réussi à obtenir quelque-chose que nous étions capables d'interpréter dans la question bonus, mais elle est présente dans le Rmarkdown